Fall 12-13-2018

From Probabilistic Socio-Economic Vulnerability to an Integrated Framework for Flash Flood Prediction

Sepideh Khajehei
Portland State University

Recommended Citation
Khajehei, Sepideh, "From Probabilistic Socio-Economic Vulnerability to an Integrated Framework for Flash Flood Prediction" (2018). *Dissertations and Theses.* Paper 4666.
https://doi.org/10.15760/etd.6550

From Probabilistic Socio-Economic Vulnerability to an Integrated Framework for Flash Flood Prediction

by

Sepideh Khajehei

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Civil and Environmental Engineering

Dissertation Committee:
Hamid Moradkhani, Chair
Scott Wells
Max Nielsen-Pincus
Robert Fountain

Portland State University
2018

Abstract

Flash flood is among the most hazardous natural disasters, and it can cause severe damages to the environment and human life. Flash floods are mainly caused by intense rainfall and due to their rapid onset (within six hours of rainfall), very limited opportunity can be left for effective response. Understanding the socio-economic characteristics involving natural hazards potential, vulnerability, and resilience is necessary to address the damages to economy and casualties from extreme natural hazards. The vulnerability to flash floods is dependent on both biophysical and socio-economic factors. This study provides a comprehensive assessment of socio-economic vulnerability to flash flood alongside a novel framework for flash flood early warning system. A socio-economic vulnerability index was developed for each state and county in the Contiguous United States (CONUS). For this purpose, extensive ensembles of social and economic variables from US Census and the Bureau of Economic Analysis were assessed. The coincidence of socio-economic vulnerability and flash flood events were investigated to diagnose the critical and non-critical regions. In addition, a data-analytic approach is developed to assess the interaction between flash flood characteristics and the hydroclimatic variables, which is then applied as the foundation of the flash flood warning system. A novel framework based on the D-vine copula quantile regression algorithm is developed to detect the most significant hydroclimatic variables that describe the flash flood magnitude and duration as response variables and estimate the conditional quantiles of the flash flood characteristics. This study can help mitigate flash flood risks and improve recovery planning, and it can be useful for reducing flash flood impacts on vulnerable regions and population.

Acknowledgments

I would like to express the deepest appreciation to my advisor, Dr. Hamid Moradkhani, who has the attitude and the substance of a genius: he continually and convincingly conveyed a spirit of adventure in regard to research and an excitement in regard to teaching. Without his guidance and persistent help, this thesis would not have been possible. I also would like to thank my committee members, Dr. Scott Wells, Dr. Max Nielsen-Pincus and Dr. Robert Fountain, for their support and for their timely review of my thesis. I greatly appreciate their willingness to serve in my thesis committee. For all the members of my research group, I thank you all for your hard work and positive attitude during all of our collaboration and providing a great friendly atmosphere in the lab.

I must thank my family for their endless support. Thank you to my parents for their generous support both financially and mentally throughout my academic journey. Special thanks to my Aunt, who has been my guardian all the time and without her financial support, it would not be possible for me to pursue my goals. Lastly, I appreciate my sister for being my best friend, who has provided me with a strong love shield that always protects me from sadness.

Table of Contents

Abstract i

Acknowledgments ii

List of Tables vi

List of Figures vii

1 Introduction 1

1.1 Flash Flood 1

1.2 Socio-Economic Vulnerability 1

1.3 Flash Flood Prediction 3

1.4 Probabilistic Techniques vs. Black Box Models 7

1.5 Objectives of Dissertation 8

2 Spatial Variation of Socio-Economic Vulnerability and Flash Flood Characteristics in the Contiguous United States 10

2.1 Background 10

2.2 Socio-Economic Vulnerability Index (SEVI) 11

2.3 Flash Flood Clustering 15

2.4 Data 17

2.4.1 Socio-Economic Data 17

2.4.2 Flash Flood Data 17

2.5 Results and Discussion 19
2.5.1 Socio-Economic Vulnerability over CONUS. 19
2.5.2 Spatial Distribution of Flash Flood Characteristics 23
2.5.3 Socio-Economic Vulnerability and Flash Flood Characteristics 28
2.5.4 Flash Flood Fatality vs. Socio-Economic Vulnerability 30
2.6 Summary and Conclusion 32
3 Assessment of the Influence of Hydroclimatic Variables on Flash Flood 34
3.1 Background 34
3.2 D-Vine Copula Based Quantile Regression Model 35
3.3 Climate Forecast System 36
3.4 Flash Flood Characteristics 38
3.5 Methodology 40
3.5.1 Pair Copula Construction 40
3.5.2 D-Vine Quantile Regression Model Construction 45
3.6 Results and Discussion 49
3.6.1 Influential Variables on Flash Flood Duration 49
3.6.2 Influential Variable on Flash Flood Magnitude 60
3.7 Summary and Conclusion 69

4 Modeling the Joint Influence of Hydroclimatic Variables on Flash Flood Characteristics using D-Vine Quantile Regression Model 71

4.1 Background 71

4.2 Methodology 72

4.3 Results and Discussion 72

4.3.1 Predictive Power of D-Vine Copula Quantile Regression Model 73

4.3.2 Flash Flood Prediction System Success Rate 77

4.3.3 Success Rate vs. Number of Predictors 82

4.3.4 Success Rate vs. Predictors 87

4.4 Summary and conclusion 93

5 Conclusion and Future Studies 95

6 References 99

List of Tables

Table 2-1. Variables used in this study to quantify socio-economic vulnerability index (SEVI). The last column indicates the overall correlation of each variable to vulnerability (i.e. a positive sign means that increase in variable will increase SEVI, and vice versa). 18

Table 2-2. Retaining components explained variances at the County Level 20

Table 2-3. Retaining Components Explained Variances at the State Level..... 22

Table 3-1. The hydroclimatic variables obtained from the CFSv2 database 40

List of Figures

Figure 2-1. The methodology employed to calculate the Socio-Economic Vulnerability Index (SEVI)..... 16

Figure 2-2. Socio-Economic Vulnerability Index (SEVI) Spatial Distribution at County Level..... 21

Figure 2-3. Socio-Economic Vulnerability Index (SEVI) spatial distribution at state level. 23

Figure 2-4. Spatial Clustering of Flash Flood Frequency over CONUS. 24

Figure 2-5. Spatial Clustering of Flash Flood magnitude over CONUS. 25

Figure 2-6. Spatial Clustering of Flash Flood Duration over CONUS. 26

Figure 2-7. Spatial Clustering of Flash Flood Severity over CONUS. 27

Figure 2-8. Flash Flood Characteristics over the County-Scale 28

Figure 2-9. Intersection of SEVI and Flash Flood Characteristics 29

Figure 2-10. Flash Flood Fatalities in comparison to Socio-Economic Vulnerability. 31

Figure 3-1. Stationarity analysis of seasonal precipitation from CFS. 38

Figure 3-2. Gridded USGS Stations..... 39

Figure 3-3. Schematic Diagram Representing Five Dimensional Copula Density with R-Vines. 43

Figure 3-4. Schematic Diagram Representing Five Dimensional Copula Density with C-Vines. 44

Figure 3-5. Five Dimensional Copula Density with C-Vines. 45

Figure 3-6. Kendall's Tau-A correlation between flash flood duration and the chosen hydroclimatic variables, precipitation amount and rate, solid moisture, runoff, storm runoff, precipitable water, and vegetation. 50

Figure 3-7. The hydroclimatic variables influencing the flash flood duration at different USGS station across the CONUS. 52

Figure 3-8. Number of predictors (i.e., covariates) chosen in eaah USGS station to construct the D-vine copula model for flash flood duration. 54

Figure 3-9. Percentage of stations dominated by a specific number of predictors (covariates) in USGS water management regions 1 to 9. 56

Figure 3-10. Percentage of stations dominated by specific number of predictors (covariates) in USGS water management regions 10 to 18. 57

Figure 3-11. Percentage of stations with specific influencing hydroclimatic variable in USGS water management regions 1 to 9. 58

Figure 3-12. Percentage of stations with specific influencing hydroclimatic variable in USGS water management regions 10 to 18. 59

Figure 3-13. Kendall's Tau-A correlation between the flash flood magnitude and the chosen hydroclimatic variable, precipitation amount and rate, solid moisture, runoff, storm runoff, precipitable water, and vegetation. 61

Figure 3-14. The hydroclimatic variables influencing the flash flood magnitude at different USGS station across the CONUS. 63

Figure 3-15. Number of predictors (i.e., covariates) chosen at the USGS stations for flash flood magnitude. 64

Figure 3-16. Percent of stations dominated by specific number of predictors in USGS water management regions 1 to 9. 65

Figure 3-17. Percent of stations dominated by specific number of predictors in USGS water management regions 10 to 18. 66

Figure 3-18. Percentage of stations dominated by specific number of predictors in USGS water management regions 1 to 9. 67

Figure 3-19. Percentage of stations dominated by specific number of predictors in USGS water management regions 10 to 18. 69

Figure 4-1. Median of Flash Flood Duration calculated for Observation (Top) and simulation (Middle), and the Absolute Difference between Observation and Modeled Flash Flood Duration (Bottom), all shown in hours. 75

Figure 4-2. Flash Flood magnitude Median calculated for Observation (Top) and Modeled (Middle), and the Absolute Difference between Observation and Modeled Flash Flood Duration (Bottom). 76

Figure 4-3. Success rate for the modeled flash flood duration (top) and magnitude (bottom) using the D-Vine Copula Regression model. 77

Figure 4-4. USGS Water Resources Regions. The colors used for each region is similarly utilized for plotting Figures 4-5 and 4-6. 79

Figure 4-5. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model for the flash flood duration and magnitude for regions 1 to 9.80

Figure 4-6. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model for the flash flood duration and magnitude for region 10 to 18. 81

Figure 4-7. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the number of predictors for the flash flood duration including regions 1 to 9..... 83

Figure 4-8. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the number of predictors for the flash flood duration including regions 10 to 18..... 84

Figure 4-9. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the number of predictors for the flash flood magnitude including regions 1 to 9..... 85

Figure 4-10. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the number of predictors for the flash flood magnitude including region 10 to 18. 86

Figure 4-11. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the predictors involved for modeling flash flood duration in regions 1 to 9. 88

Figure 4-12. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the predictors for the flash flood duration including region 10 to 18. 89

Figure 4-13. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the predictors for the flash flood magnitude including region 1 to 9. 91

Figure 4-14. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the predictors for the flash flood magnitude including region 10 to 18. 92

1 Introduction

1.1 Flash Flood

Flash floods cause extensive damage and disruption to societies and environment, and they are among the deadliest natural hazards worldwide. Several studies have assessed the impacts of flash floods events around the world with regards to substantial financial losses, destruction of infrastructure, displacement, and fatalities (Ashley and Ashley 2008; Armah et al. 2010; Kotzee and Reyers 2016). The flash flood hazard is expected to increase in frequency and severity as a result of the global climate change, severe weather in the form of heavy rains, and river discharge conditions (Kleinen and Petschel-Held 2007; Borga et al. 2011). There is an extensive body of research on the effects of development in riverine floodplains, whereas less attention has been given to flash flooding and flash flooding impacts (Hapuarachchi et al. 2011; Cutter et al. 2017).

1.2 Socio-Economic Vulnerability

Although significant advances have been made in sustainability and vulnerability science, especially the conceptualization and representation of vulnerability within the human-environment systems (Adger 2006; Folke 2006; Cutter and Finch 2008; Cutter 2012), there are still differences in interpretation of vulnerability between the risk/hazards studies and human/environmental research communities. However, both communities acknowledge that the composition of vulnerability is driven by exposure, sensitivity, and response or resilience to the compound effect of both environmental and social systems (Turner et al. 2003; Kasperson 2005; Blaikie et al. 2014).

Social vulnerability is a measure of both the sensitivity of a population to natural hazards and its ability to respond to and recover from the impact of hazards (Shirley et al. 2012). It is a multivariate metric, one that is not easily captured with a single variable. There is ample field-based evidence from population and social groups for understanding the societies' characteristics and investigating what makes them more sensitive to the effects of natural hazards and what reduces their ability to adequately respond and recover from a natural hazard (Center 2002; Council 2006).

A region with advanced socio-economic status is generally less vulnerable to disasters, and its coping mechanisms are more advanced and efficient (Klein et al. 2003). Indicators of social and economic status can include per capita income, percentage and extent of access to public amenities, and many more socio-economic variables (Hahn et al. 2009; Malakar and Mishra 2017). To this end, an index-based based vulnerability assessment is a comprehensive tool that helps comparing and ranking areas in terms of their vulnerability (Cutter et al. 2003; Kelkar et al. 2011; Ahmadalipour and Moradkhani 2018).

Physical science and engineering advancements are essential for coping with flash floods, particularly as hydrologists and meteorologists strive to understand the factors that allow predicting flash floods. However, such advancements will only make a difference if the recognition and understanding of warnings, warning response, and risk communication are also improved. Exposure to flash flooding and vulnerability to loss continue to increase, even as the ability to forecast events increases and early warning systems are implemented (Montz and Gruntfest 2002).

Špitalar et al. (2014) studied flash flood fatalities and injuries from 2006 to 2012 in the United States. An interesting result from this study was that most human-impacting events occur in rural settings. However, when a flash flood occurs in an urban area, there are many more human impacts (i.e., injuries and fatalities) per event. Notably, there is a likelihood that the frequency and number of flood fatalities increase over the U.S. in the coming decades due to the following reasons. First, the U.S. population continues to urbanize (Economic 2007); urbanized basins respond quickly to rainfall due to reduced infiltration and faster conveyance of water in channelized canals. This urbanization process puts an increasing number of individuals and their properties in flood-exposed environments (Montz and Gruntfest 2002). Second, climate change studies have projected an intensifying hydrologic cycle under future emission scenarios, resulting in more intense rainfall events and exacerbated flash floods (Trenberth et al. 2003). However, very few studies have evaluated the socio-economic vulnerability of the United States to flash flood events. Such a study can help in the advancement of flash flood risk mitigation planning as well as socio-hydrologic management to reduce the trend of flash flood fatalities in the United States and beyond.

1.3 Flash Flood Prediction

Identifying the occurrence of flash flooding, estimating the associated risk, and implanting effective mitigation measures require high spatial resolution flash flood forecasting with adequate lead-time. Several types of models have been used to forecast flash flooding, including physical simulation models.

The occurrence of flash flood is linked, on one side, to the size of the concerned catchment, and on the other side, to the activation of surface runoff that become the prevailing transfer process (Marchi et al. 2010). Moreover, surface runoff may be affected by different processes, owing to the combination of intense rainfall, soil moisture, and soil hydraulic properties as well as the land use alteration, urbanization and vegetation (Marchi et al. 2010; Du et al. 2012; Hong et al. 2013; Miller et al. 2014; Binley et al. 2015; Verhoest et al. 2015; Gourley et al. 2017; Saharia et al. 2017b).

Flash floods are influenced by meteorological hazards compounded by hydrologic and human behavioral setting. There can be situations where an intense rainfall event yields no flash flooding in an unpopulated area with dry, sandy soils. Conversely, seemingly small rainfall amounts occurring over urban areas with impervious surfaces can cause catastrophic consequences (Gourley et al. 2017). Therefore, beside improving the flash flood forecasting and warning systems, there is a need for an extensive assessment of the hydroclimatic and environmental characteristics that are involved in emerging flash flood events.

Gourley et al. (2017) introduced the Flooded Locations and Simulated Hydrographs (FLASH) project that provides a suite of products to advance the state of science in flash flood prediction. The FLASH project is an outgrowth from the Multi- Radar Multi-Sensor (MRMS) system that generates a suite of severe weather, aviation, and hydrometeorological products for the NWS across the conterminous United States (CONUS) and southern Canada (Zhang et al. 2016). The FLASH project focuses on forecasting and quantifying the flash flood magnitude. However, there is a need to

understand the interaction between hydrometeorological variables in the process of shaping flash floods.

The current state of flash flood warning systems is based on the use of models that simulate hydrological processes at watershed scale. Two important variables that control runoff generation and serve as input to models are initial soil moisture and rainfall distribution (Wanders et al. 2014; Najafi and Moradkhani 2015). In addition, the high amount of precipitable water may lead to an increase in the highest possible rainfall intensities and an increase in the frequency of extremely high daily rainfall totals, regardless of how average rainfall may change (Halmstad et al. 2013; Reza Najafi and Moradkhani 2013). A consequence of higher rainfall rates in a warmer world has increased flash flooding and riverine flooding (Kunkel et al. 2013; Zarekarizi et al. 2018).. Furthermore, vegetation can help slow down runoff and prevent flooding. Vegetation and specifically trees reduce the risk of flooding. The vegetation roots hold the soil together, and during the growing season they absorb water. When there is a lack of vegetation, however, there is not much to stop the water from running off (Hapuarachchi et al. 2011; Goulden and Bales 2014; Vanuytrecht et al. 2014).

Over the past decades, there have been an increasing pursuit to improve the biophysical flood forecast simulation models in different parts of the world (Najafi and Moradkhani 2014; Douinot et al. 2016; Kauffeldt et al. 2016; Yan and Moradkhani 2016; Wang et al. 2017). Recent advances in flash flood forecasting include new techniques for developing Quantitative Precipitation Forecasts (QPFs) and Quantitative Precipitation

Estimates (QPEs), employing remotely sensed precipitation data, flow forecast models, and forecast uncertainty estimates.

Flow comparison, rainfall comparison, and flash flood susceptibility assessments are among the popular techniques for predicting flash flood occurrence (Smith and Austin 2000; Collier and Fox 2003; Reed et al. 2007). The flow comparison methods are shaped based on comparing the modelled flow values and the long-term simulated records. Therefore, flow comparison is capable of accurately predicting the specific flash flood stage and flow if there is a combination of high-quality QPF and perfectly calibrated physically based distributed hydrological model (Hapuarachchi et al. 2011).

On the other hand, the rainfall comparison or flash flood guidance methods compare the rainfall required over a specific area to generate flooding flow at its outlet with the rainfall forecast. Therefore, these techniques are useful tools for flash flood warning and estimating the hazard at different time frames (Norbiato et al. 2009). However, they are not as accurate as flow comparison techniques for estimating the flash flood magnitude.

The flash flood susceptibility assessment techniques evaluate the physical and environmental vulnerability of a certain catchment to flooding due to extreme rainfalls. The flash flood susceptibility assessment procedures are operationally simple, although the forecasts may be less reliable since the method is essentially interrogative (Hapuarachchi et al. 2011).

The flash flood forecasting systems described above are employing simplified physical equations, which boost the uncertainty and deteriorate the accuracy in model

results. The physical models have two main shortcoming. First, they are computationally expensive and time consuming, whereas the simplified equations results in a certain level of inaccuracy in the results. In contrast, data-driven models are capable of using historical relationships between different variables such as hydroclimatic parameters and flash flood characteristics in order to forecast future flash flooding characteristics. The data-driven models include statistically-based or machine learning methods based on artificial intelligence tools (Hong 2008; Furquim et al. 2014). Using such techniques will overcome the pitfalls from the physical models for simulating complex phenomena (Toukourou et al. 2009).

1.4 Probabilistic Techniques vs. Black Box Models

Data-driven models can be categorized into black box and statistical models. Black box models such as neural networks and machine learning techniques have limited ability to explicitly identify possible causal relationships. In contrary, statistical models such as regression are superior to neural networks for identifying possible causal relationships (Tu 1996). For instance, in statistical models, the model developers have the advantage of determining the variables that are most strongly predictive of an outcome. In addition, through stepwise variable selection process, a number of independent variables that are not related to a particular outcome can be eliminated by employing statistical procedures. Since one of the main purposes of this study is to identify the driving variables for flash flooding, the statistical models are the superior candidate to black box models (Eftekhar et al. 2005; Yilmaz 2010).

Furthermore, the black box models requires greater computational resources comparing to the statistical models, and for developing an operational warning system the computation cost of the model plays a vital role (Kolman and Margaliot 2005). Overall, the statistical models remain the clear choice, since the primary goal of this model development is to investigate possible causal relationships between the hydro-climatic variables and flash flooding.

1.5 Objectives of Dissertation

The primary objectives of this study are summarized as follows:

1. Perform a comprehensive and multi-dimensional assessment of socio-economic vulnerability and its coincidence with flash flood characteristics over the contiguous United States (CONUS).
2. Evaluate the socio-economic vulnerability at both the county and state level and detect their differences.
3. Assess the after event impacts by comparing the flash flood fatalities through out the CONUS.
4. Identify the resemblance and heterogeneity of flash flood spatial clustering and socio-economic vulnerability over the CONUS.
5. Develop a data analytic approach to assess the interaction between the flash flood characteristics and the hydroclimatic variables over the CONUS.
6. Identify the influencing variables on flash flood magnitude and duration at different regions across the CONUS.

7. Introduce a multivariate statistical flash flood forecasting system using D-Vine Copula quantile regression model which is capable of estimating uncertainty.

2 Spatial Variation of Socio-Economic Vulnerability and Flash Flood Characteristics in the Contiguous United States

2.1 Background

The compound effect of population growth, inappropriate land-use planning, and environmental degradation together with the effect of global climate change is the main driver of the increasing losses caused by the impact of natural hazards (Aroca-Jiménez et al. 2018). Flash floods are identified as one of the natural hazards with the highest capacity to generate risk, including both the socio-economic and human impact on a global scale (Terti et al. 2015). The high risk associated to flash flood is due to the main characteristic of flash flood that is associated with rapid onset happening in a relatively short time ; therefore, significantly reduces the warning and response time of the population and relevant agencies (Javelle et al. 2016).

As a result, it is critical to develop an indicator of the disaster risk and vulnerability to flash flood. Such an indicator should be able to help the decision makers to assess the potential impact of natural hazards and at the same time identify the most vulnerable social group and areas (Birkmann et al. 2013).

Evaluation of socio-economic vulnerability usually consists of the construction of vulnerability indices representing the inherent characteristics or qualities of social systems that create the potential impact and the static nature of society (Cutter and Finch 2008; Cutter 2012; Terti et al. 2015). However, very few studies have evaluated the socio-economic vulnerability of the United States to flash flood events (Cutter et al. 2013). Such studies can help advance flash flood mitigation planning as well as socio-hydrologic

management to reduce flash flood fatalities in the United States and beyond. For instance, by understanding the vulnerability of a region to flash flooding, it will be possible to assign various evacuation policies for different regions according to their vulnerability. A region (i.e. county) with high levels of vulnerability should be asked to evacuate at lower flash flood hazards, compared to the regions that have low vulnerability. In addition, special programs can be initiated to assist the regions that are associated with high hazard (frequent and intense flash floods) as well as high vulnerability.

The current study is among the first comprehensive and multi-dimensional assessment of socio-economic vulnerability and its coincidence with flash flood characteristics over the contiguous United States (CONUS). This study integrates the socio-economic vulnerability with the flash flood characteristics at both the county and state level, and it explores the spatial distribution of flash flooding across the CONUS. The results will determine the resemblance and heterogeneity of flash flood spatial clustering and vulnerability of regions over the CONUS.

2.2 Socio-Economic Vulnerability Index (SEVI)

Cutter et al. (2003) introduced the Social Vulnerability Index (SoVI) to examine the spatial patterns of social vulnerability to natural hazards at the county level in the United States. The SoVI was developed by finding the social characteristics consistently identified within the research literature as contributing to vulnerability. These variables should demonstrate the socioeconomic status of the given region as well as its capacity for recovery from the impacts of natural hazards. Selection of the specific variables to represent the socio-economic vulnerability is specific to the study purpose. However, the

most common characteristics include demographic socioeconomic status, the quality of human settlements and the built environment (Cutter et al. 2003). Due to the high number of variables describing the social vulnerability, statistical procedures are commonly used to reduce data dimensions. The Principal Component Analysis (PCA) (Manly and Alberto 2016) has been widely utilized for decreasing the dimension of data in order to create a single and consolidated index of social variables (Filmer and Pritchett 1998; Cutter et al. 2003; McKenzie 2005; Vyas and Kumaranayake 2006; Cutter and Finch 2008; Wigtil et al. 2016). Based on this procedure, the component which explains the majority of the variance of data will be chosen to calculate the SEVI. The most important step in constructing the index is scaling the chosen components. Each component represents a different element of vulnerability, therefore positive values raise the vulnerability index and negative values reduce it (Schmidtlein et al. 2008). One shortcoming of using PCA-based index is that this procedure falls short in dealing with missing values. To overcome this problem, some studies have suggested replacing the missing values with a value of zero (Cutter et al. 2003; Cutter and Finch 2008; Wigtil et al. 2016). However, a zero value cannot accurately represent the true vulnerability based on the particular variable of interest, and it would underestimate the level of vulnerability for the affected regions.

This study introduces a new algorithm for building the socio-economic vulnerability, which is based on Probabilistic Principal Components Analysis (PPCA) introduced by Tipping and Bishop (1999) and overcomes the shortcomings of PCA. The PPCA is a probabilistic formulation of PCA based on a Gaussian latent variable model, which has the capability to estimate missing values, and therefore, there is no need for

replacing the missing values with zero. PPCA retains the characteristics of PCA such as the principle scores and loadings. The expectation maximization (EM) algorithm is used to estimate the parameters of PPCA model, which will consequently allow the framework to deal with the missing values (Nasrabadi 2007; Nyamundanda et al. 2010).

To generate the SEVI, the social and economic variables were chosen based on the previous studies and they are presented in Data section. These variables include demographic socioeconomic status, race and ethnicity, age, employment and gender, housing and transportation, and industrial economy. The demographic socioeconomic status demonstrates the ability of the society to absorb losses and enhance resilience to hazard impacts. In addition, wealth enables communities to absorb and recover from losses more quickly due to insurance, and social safety nets. Race and ethnicity impose language and cultural barriers that affect access to post-disaster funding and residential locations in high hazard areas. Extreme of age spectrum affect the movement. For example, parents lose time and money caring for children when daycare facilities are affected. The potential loss of employment following a disaster exacerbates the number of unemployed workers in a community, contributing to a slower recovery from the disaster. Loss of sewers, bridges, water, communications, and transportation infrastructure compounds potential disaster losses. The loss of infrastructure may place an insurmountable financial burden on smaller communities that lack the financial resources to rebuild. The income from the industries provides an indicator of the state of economic health of the community, and longer term issues with recovery after an event.

All the input variables should be normalized, therefore the z-scores were calculated for all the data with zero mean and unit standard deviation. The PPCA was performed with the normalized input variables. PPCA returns a set of orthogonal components that are the linear combinations of the original variables. In this step, it is necessary to choose the number of components representing the chosen variables. In this study, the Kaiser criterion was used to select the number of components, which retains the components with eigenvalues greater than 1. Then, the varimax rotation was applied to the retaining components in order to reduce the number of highly loading variables on each component. Before formulating the SEVI, the resulting components should be scaled based on their influence on the socio-economic vulnerability. The influence of the chosen variables are presented in Table 2-1. Furthermore, the loadings corresponding to each variable will be evaluated to find the highly influencing variable on each components. Afterwards, if a component indicates low socio-economic vulnerability, the corresponding component scores are multiplied by -1, and vice versa. Followed by the suggestion of Schmidtlein et al.(2008), the component scores were combined by weighted sum using explainable variance. Finally, the resulting score was normalized between 0 and 1 to calculate the final SEVI. Figure 2-1 presents the schematic for SEVI development procedure in this study.

The methodology was applied to the socio-economic variables at county and state levels to compare the vulnerability at different spatial scales. For clear presentation, the regional units (i.e. state and county) were divided into three categories; low vulnerable which are the regions placed in the lower 10^{th} percentile quantile, high vulnerable which are the ones that have a SEVI higher than the 90^{th} percentile, and the remaining are

considered medium vulnerable. The calculated socio-vulnerability is a unitless spatial measure, and it is designed to use as a comparative value across geographic locations.

2.3 Flash Flood Clustering

To map the SEVI and flash flood characteristics and their spatial analogy, the United States Geological Survey (USGS) stations were spatially clustered based on each flash flood characteristics (i.e., magnitude, duration, frequency, and severity). This study employed the Getis-Ord hotspot analysis or this purpose. The hotspot analysis uses the flash flood characteristics to identify the locations of statistically significant hot spots and cold spots in the USGS data.

A high Z score and small P value for a region indicates a significant hot spot (i.e. significant clusters of high values). A low negative Z score with a high absolute value and small P value indicates a significant cold spot (i.e. significant clusters of low values). In simple words, hotspots are the regions that the variable of interest (e.g. flash flood magnitude or frequency) is high and the surrounding regions also indicate large values. Therefore, they are associated with higher hazard. The higher the Z score (absolute value) the denser is the clustering. A Z score near zero means no spatial clustering.

The Getis-Ord local statistic is given as:

$$G_I^* = \frac{\sum_{J=1}^{n} w_{i,j}x_j - \bar{X}\sum_{J=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{J=1}^{n} w_{i,j}^2 - (\sum_{J=1}^{n} w_{i,j})^2\right]}{n-1}}} \quad \textbf{(2-1)}$$

Where x_j is the attribute value for featurej, $w_{i,j}$ is the spatial weight between features i and j, and n is the total number of features and:

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \tag{2-2}$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{X})^2} \tag{2-3}$$

The G_I^* statistic is a z score, so no further calculations are required.

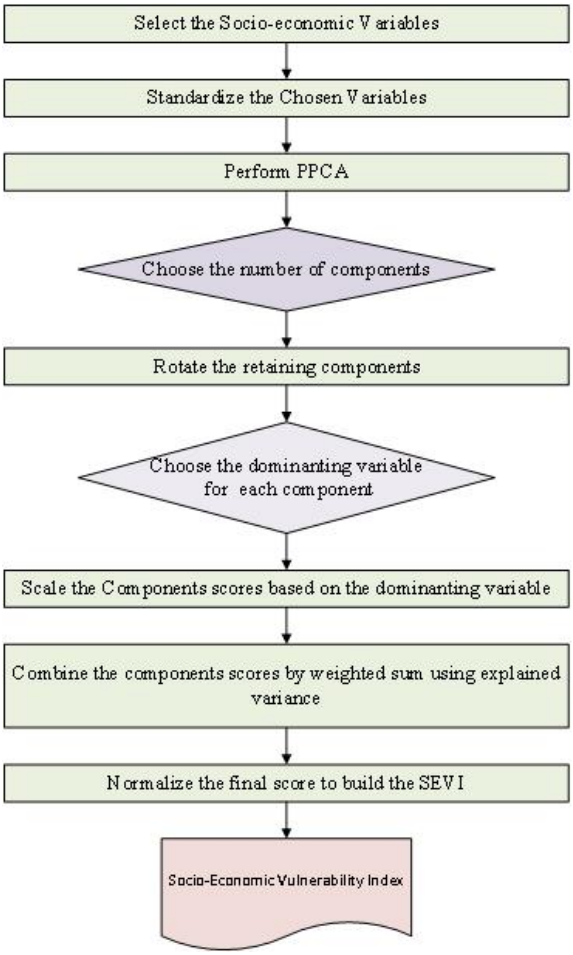

Figure 2-1. The methodology employed to calculate the Socio-Economic Vulnerability Index (SEVI).

2.4 Data

2.4.1 Socio-Economic Data

The categories that develop the basis of the Socio-Economic Vulnerability index (SEVI) are identified as: 1) Demographic socioeconomic status, 2) Race and ethnicity, 3) Age, 4) Employment and gender, 5) Housing and transportation, and 6) Industrial economy. The data from the first five categories are named Social variables and are acquired from the 2015 American Community Survey (ACS) 5-year Estimates (https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml). The industrial economy variables are collected from the Bureau of Economic Analysis (https://www.bea.gov/iTable/index_regional.cfm). Descriptions of the chosen variable are summarized in Table 2-1.

2.4.2 Flash Flood Data

The Hydrometeorology and Remote Sensing (HyDROS) group at the university of Oklahoma created a data-base of flooding information. The Unified Flash Flood Database is collected from a variety of sources such as gauge measurements of streamflow by United States Geological Survey (USGS), flash flooding reports in the National Weather Service (NWS) Storm Events Database, and public survey responses on flash flood impacts collected during the Severe Hazards Analysis and Verification Experiment (SHAVE) (Gourley et al. 2010, 2017). The desirable spatial coverage of NWS reports, and automated data collection mechanism of USGS streamflow records makes the SHAVE dataset one of the most representative flash flood databases in the United States. It is publicly available for free (https://blog.nssl.noaa.gov/flash/database/).

Table 2-1. Variables used in this study to quantify socio-economic vulnerability index (SEVI). The last column indicates the overall correlation of each variable to vulnerability (i.e. a positive sign means that increase in variable will increase SEVI, and vice versa).

Categories	Variables	Influence on the Vulnerability
Demographic Socioeconomic Status	Poverty	+
	Per capita income	-
	Median household value	-
	Percentage of population aged 25 years or older with less than 12th grade education	+
	Percentage of households receiving social security	+
	Median gross rent	-
	Percentage employment in extractive industries	+
	Percentage of households earning greater than US $200,000 annually	-
	Percentage employment in service industry	+
	Percentage civilian unemployment	+
Race and Ethnicity	Percentage Asian	+
	Percentage Black or African American	+
	Percentage speaking English as a second language with limited English proficiency	+
	Percentage Hispanic	+
	Percentage Native American	+
Age	Median age	-
	Percentage of population under 5 years or 65 and older	+
	Percentage of population under 18 years old	+
Gender	Percentage female	+
	Percentage female participation in labor force	+
	Percentage female-headed households	+
Housing and Transportation	People per unit	+
	Percentage mobile homes	+
	Percentage of housing units with no cars	+
	Percentage of population living in nursing and skilled-nursing facilities	+
	Percentage renters	+
	Percentage unoccupied housing units	-
Industrial Economy	Private industries	-
	Agriculture, forestry, fishing, and hunting	+
	Transportation and food service	-
	Accommodation and food service	-
	Governmental	-

This study used the USGS automated streamflow measurements. USGS collects instantaneous streamflow data at intervals ranging from 5 to 60 min for 10,106 gauges in the database. The NWS coordinates with local stakeholders and the USGS to define stages corresponding to action stage as well as minor, moderate, and major flooding for 3,490 stream gauge locations. Flood events are defined when streamflow exceeds the defined action stage for that gauge. There must be a 24-h difference between when streamflow drops below action stage to the next rise for it to be accounted as a separate event (Saharia et al. 2017a). This study focuses on four characteristics of flash flood at each station including frequency, magnitude, duration, and severity. Severity is defined as the magnitude divided by the duration.

2.5 Results and Discussion

The results of socio-economic vulnerability and its spatial relationship with flash flood characteristic are divided into two sections. The first section investigates the characteristics of SEVI for two US census geographical units: states and counties. Then, the resemblance of the flash flood characteristics and the SEVI is evaluated over the CONUS.

2.5.1 Socio-Economic Vulnerability over CONUS.

2.5.1.1 SEVI at County Level

The SEVI is calculated by feeding the PPCA with 32 social and economic variables described in section 2.4.1. The PPCA analysis of socio economic vulnerability resulted in retaining nine components, which explain 78% of the total variance among the U.S.

counties. Each of the components explains between 3 to 17 percent of the total variance. The dominant categories for each retaining component (i.e., the variables with the highest value of loading) are presented in Table 2-2.

Table 2-2. Retaining components explained variances at the County Level

Component	Category	% Variance Explained
1	Demographic Socioeconomic Status	16.15
2	Industrial Economy	14.27
3	Race and Ethnicity	12.13
4	Demographic Socioeconomic Status	10.46
5	Housing and Transportation	9.87
6	Demographic Socioeconomic Status	6.52
7	Age	5.32
8	Gender	3.49

Results of Table 2-2 indicate that the socioeconomic status dominates the SEVI with more than 30% of the variance explained. This finding is in agreement with the results from Cutter and Finch (2008), in which the socioeconomic status explained approximately 20% of the variance in five decades at a county level.

The calculated SEVI of each county is classified into three classes of high vulnerable, medium vulnerable, and low vulnerable. The counties with SEVI in the lower 10^{th} percentile and top 90^{th} percentile are categorized into low and high classes, respectively. Figure 2-2 shows the SEVI at the county level.

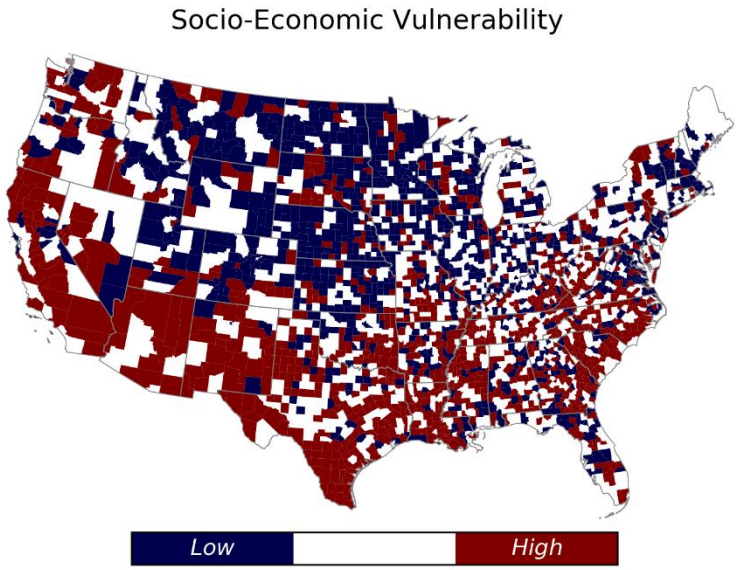

Figure 2-2. Socio-Economic Vulnerability Index (SEVI) Spatial Distribution at County Level.

As shown in Figure 2-2, the most socially vulnerable counties are concentrated in the southwest and the southern Plains (i.e, Texas and Lousiana) along the U.S.–Mexico border regions of Texas. The least vulnerable counties are located in New England and the upper Great Lakes. Results from the (Cutter et al. 2003; Cutter and Finch 2008; Wigtil et al. 2016) indicate similar spatial geographical patterns. Cutter and Finch (2008) showed that the vulnerability is spreading in the southwest region toward the U.S.–Mexico border regions of Texas in the last four decades, which can be due to the clustering of high vulnerable counties in that region, based on SEVI, calculated in this study.

2.5.1.2 SEVI at State Level

The SEVI at the State level is calculated with similar variables and techniques to the US. counties. The PPCA results are summarized in Table 2-3 with the retaining components explaining 81% of the variance. The industrial economy dominates the main component (i.e. PC1), however the socioeconomic status explains the majority of variance in the data.

Table 2-3. Retaining Components Explained Variances at the State Level

Component	Category	% Variance Explained
1	Industrial Economy	17.35
2	Demographic Socioeconomic Status	15.98
3	Housing and Transportation	12.65
4	Gender	11.49
5	Demographic Socioeconomic Status	9.38
6	Race and Ethnicity	8.74
7	Age	6.21

Figure 2-3 represent the SEVI calculated at the state level. High vulnerable states are clustered in the Great Plains and central U.S. Evaluating the SEVI and industrial economy; it is shown that the states with higher wealth indicate lower vulnerability. This is reasonable since the industrial economy is the dominant component with the highest weighting value.

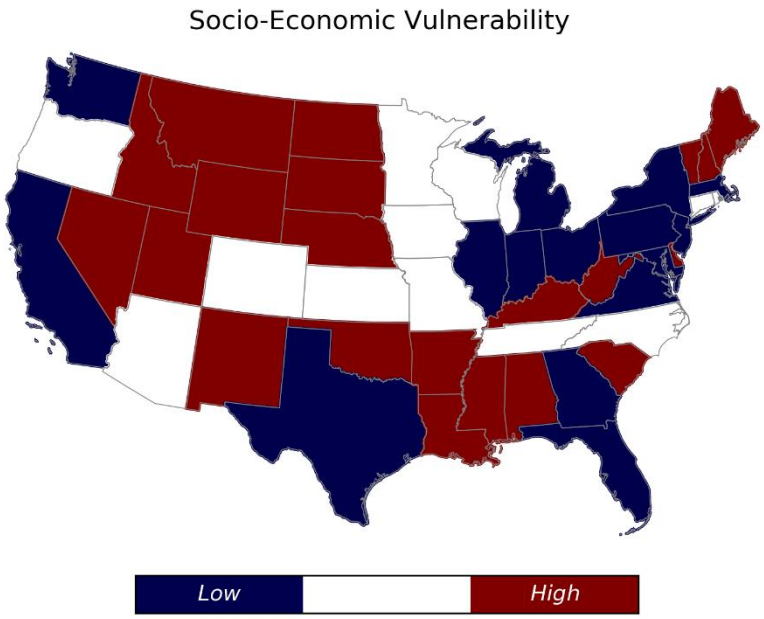

Figure 2-3. Socio-Economic Vulnerability Index (SEVI) spatial distribution at state level.

Comparing the SEVI calculated for U.S. counties (Figure 2-2) and the one for U.S. states (Figure 2-3) shows that in general, the counties and states located in the Great planes show high vulnerability. However, for states such as Texas, Florida, and California, it does not follow the same pattern. In detail, large clusters of high vulnerable counties are located in these three states. Comparing the results in Table 2-2 and 2-3 indicates that the main characteristic difference between the county and state SEVI is how the industrial economy dominates the vulnerability index. This observation may prove that the industrial money is not distributed equally among the counties in a given states.

2.5.2 Spatial Distribution of Flash Flood Characteristics

This study targets four characteristics of flash flood including frequency, magnitude, duration, and severity. Results from the hotspot analysis are summarized in

Figure 2-4 through Figure 2-7. Hotspots are representing the spatial clustering of high values for that given characteristics and spatial clustering of low values are shown by the coldspots.

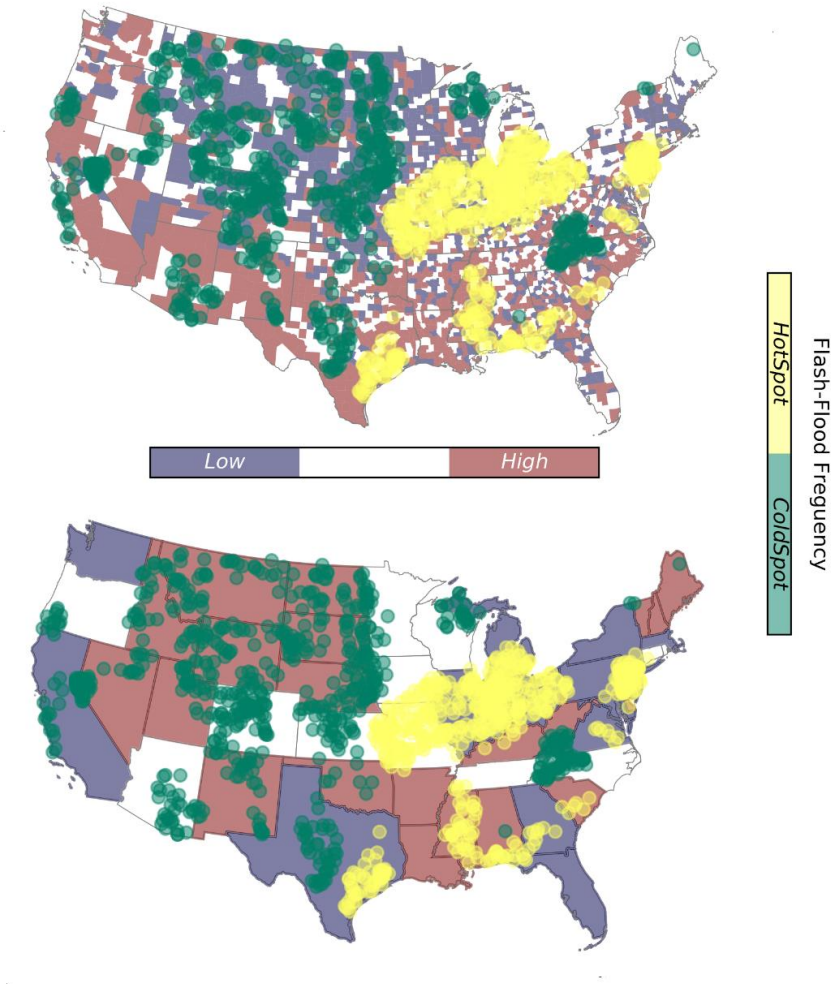

Figure 2-4. Spatial Clustering of Flash Flood Frequency over CONUS.

As shown in Figure 2-4, the Midwest experiences the largest values of flash flood frequency, whereas the cold spots are located in the Great Plains region. Eastern and southern parts of the U.S show smaller hotspot clusters and the scattered cold spots are visible in the southwest and pacific regions.

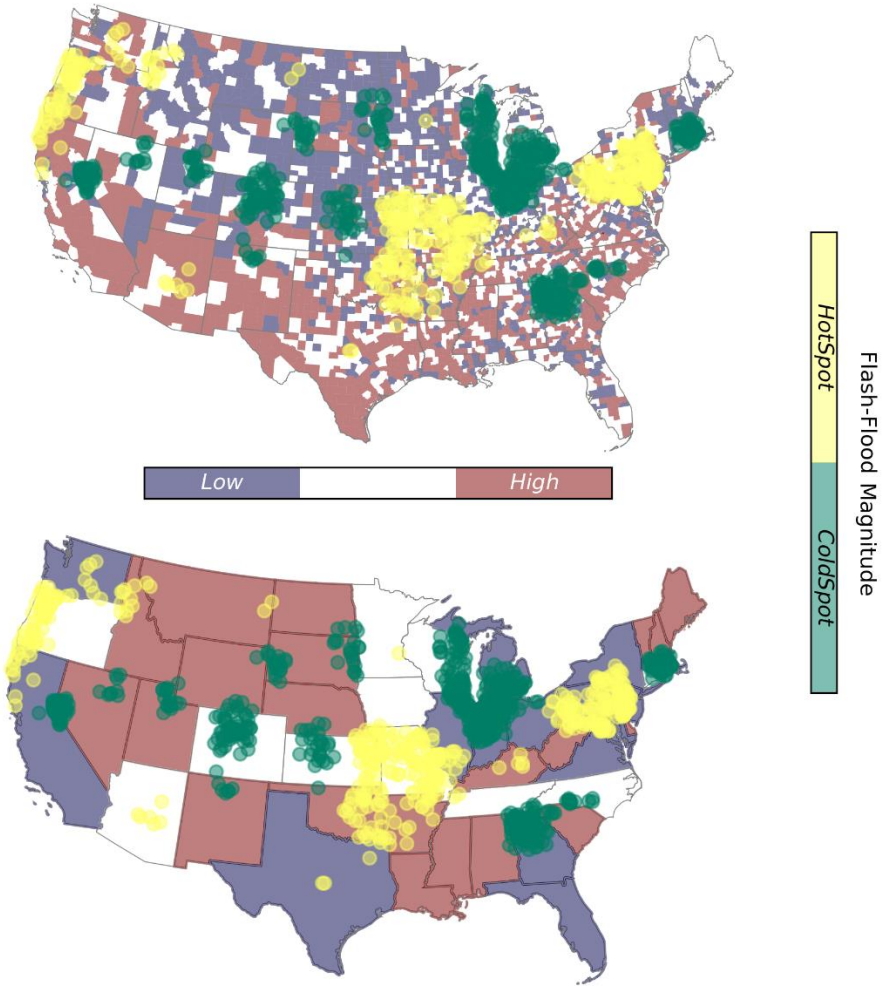

Figure 2-5. Spatial Clustering of Flash Flood magnitude over CONUS.

The flash flood magnitude has the smallest clusters comparing to rest of characteristics (Figure 2-5). The hotspot cluster is located in the Missouri Valley, East coast, and Oregon. The largest coldspot is located in the Upper North region. This indicates that the magnitude of the flash flood varies substantially throughout the CONUS.

Figure 2-6 represents the results for flash flood duration. A considerably large coldspot cluster is located in the Northeastern and part of Southeastern region. The hotspot

clusters are spread out in the lower part of Southeast, Midwest, and Great Plains. Disperse clusters of short flash flood duration are located in the Central U.S., Southwest, and pacific region.

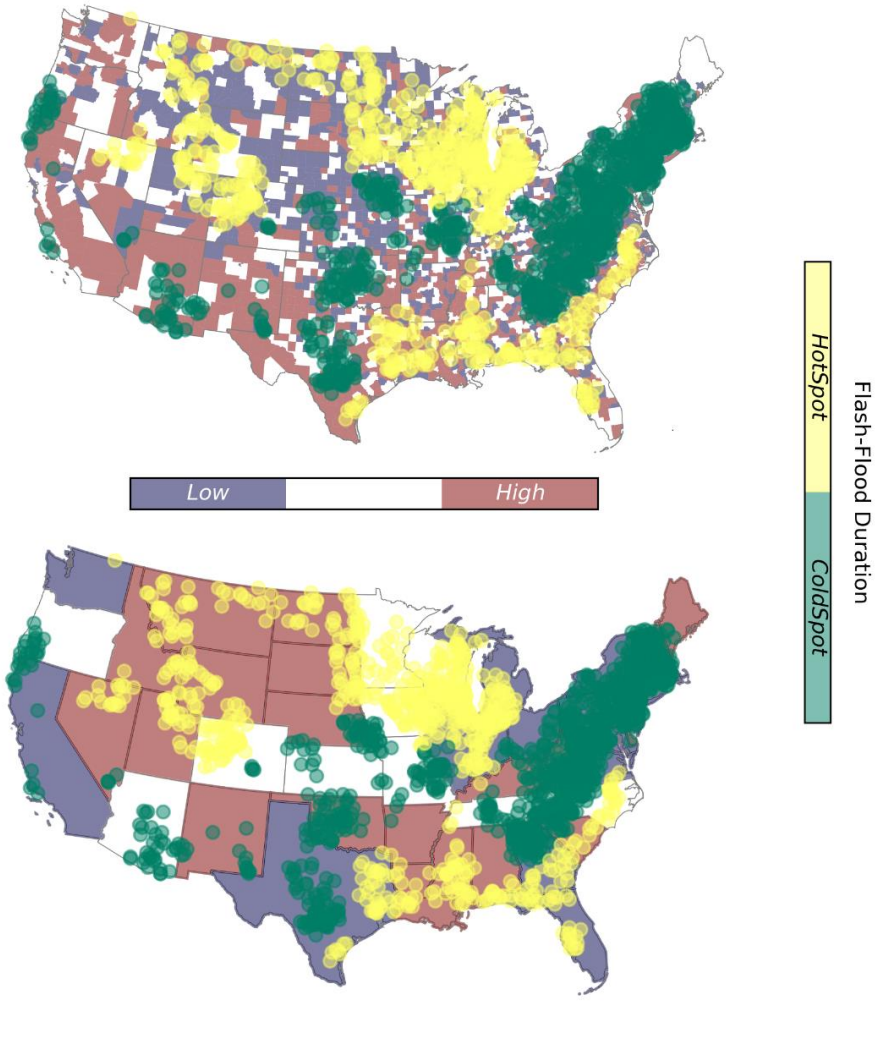

Figure 2-6. Spatial Clustering of Flash Flood Duration over CONUS.

Similar to the magnitude, the severity has small number of clusters, as it depends on the magnitude (Figure 2-7). The largest cluster is located in the Appalachians, Texas, Arizona, and West coast. On the other hand, the Missouri Valley is where the coldspot of

flash flood duration is located. This spatial pattern is observed in the study that was conducted by Saharia (2017b), and they looked at the flash flood severity over the United States by introducing a new index called flashiness.

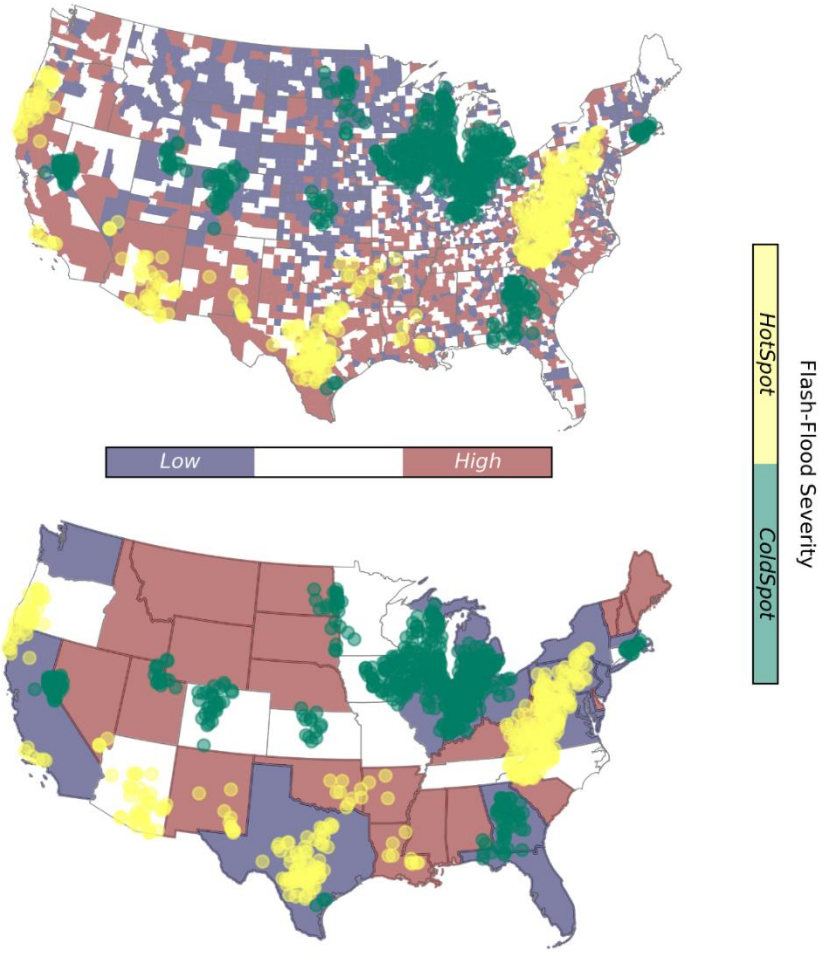

Figure 2-7. Spatial Clustering of Flash Flood Severity over CONUS.

One of the porpuses of this study is to invesigate the coincide of flash flood chararacteristics and socio-economic vulnerability over the CONUS. Therefore, the flash flood characteristis needs to be converted to county-scale from the station data. Figure 2-8 represents the flash Flood characteristics on the county level. The orange counties are the

ones where the stations show hotspot for that specific flash flood characteristics and the blue color indicates the coldspot.

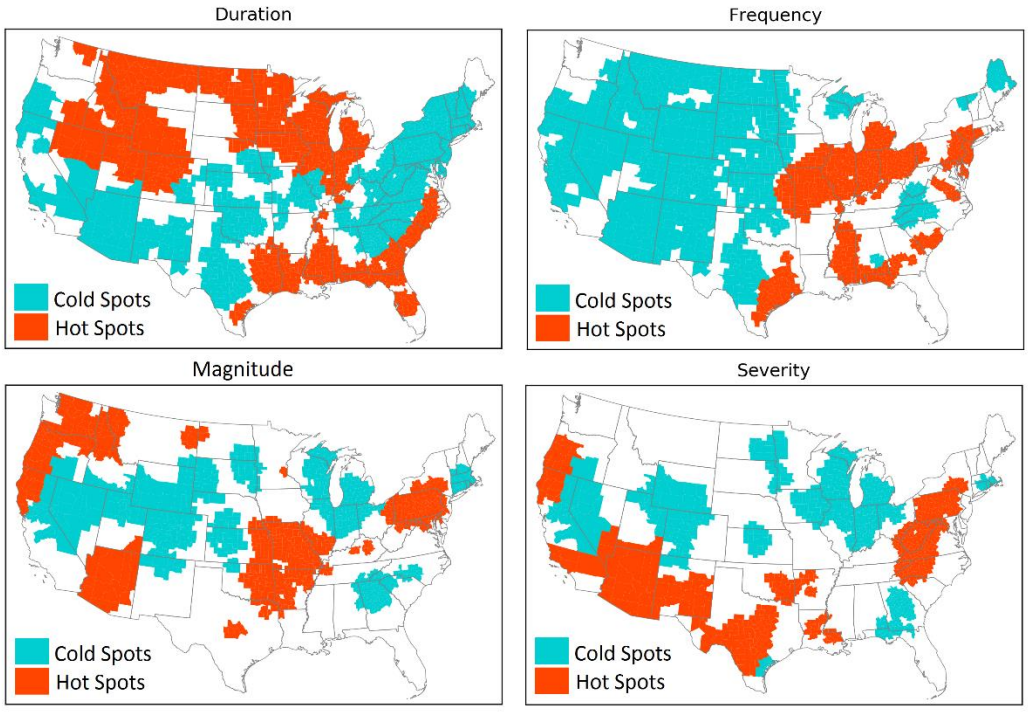

Figure 2-8. Flash Flood Characteristics over the County-Scale

The counties that are located in the Pacific Northwest region indicate hot spots for the magnitude and severity, whereas it is not the case for the duration and frequency. The majority of counties in the Northern Great Plains are located in the hotspots for flash flood duration and cold spot for frequency. The western and central couties experience less frequent flash flood comparing to the Southeast and Northern U.S. The Southwest and Southern Great Plains are receiving more severe flash floods compared to the duration.

2.5.3 Socio-Economic Vulnerability and Flash Flood Characteristics

A two-way cross tabulation is used to map the coincidence of SEVI and flash flood characteristics presented in Figure 2-9. Comparing the results with respect to flash flood

duration and frequency, the high-vulnerable-hotspot counties are located in lower Southeast. Whereas, the high-vulnerable-hotspot counties are located in the Southwest and Pacific Northwest.

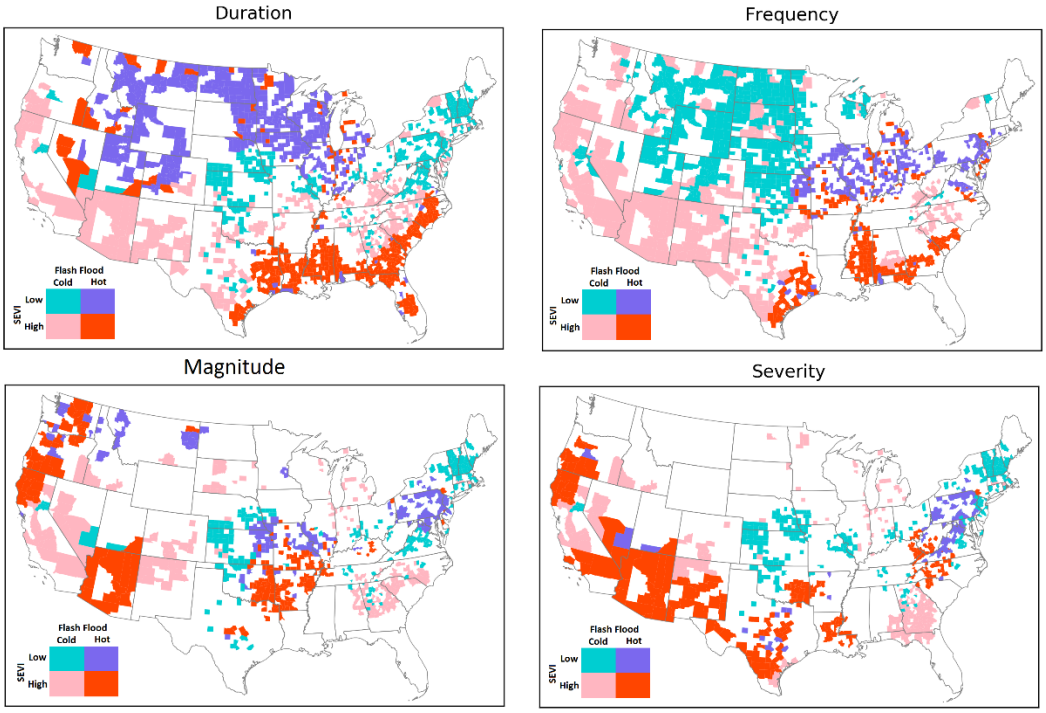

Figure 2-9. Intersection of SEVI and Flash Flood Characteristics

From Figure 2-9, the majority of high-vulnerable-coldspot (i.e., the regions associated with high vulnerability and low hazard) are located in the Southwest for duration, frequency and magnitude, however; they are scattered in regards to severity. The low-vulnerable-coldspot counties are very dense in the Northern Great Plains and Midwest, comparing the frequency with other flash flood characteristics. The low-vulnerable-hotspot are concentrated in the Northern Great Plains observing the flash flood duration. In case of a large flash flood event with vast spatial impact, the hotspot-high vulnerable regions (shown in dark red) will probably be affected the most, and since they indicate cluster of

areas that will probably be highly impacted, they are unable to receive immediate aid from adjacent counties or states either. Therefore, it is crucial to pay immediate federal attention to these regions in the case of an extensive flash flood event.

Furthermore, critical and non-critical counties are neighboring, specifically in the southern region. These counties require further assessment, whereas they might have the potential to change to the opposite status. For instance, two of the critical counties for flash flood magnitude are located in Texas adjacent to non-critical ones. A more in-depth assessment is vital to identify the real time changes in these counties.

This study aimed to generate a comprehensive SEVI, however it may still have some limitations. Although social vulnerability indices can efficiently describe broad-scale vulnerability, they may fail to capture more localized information related to exposure, sensitivity, and adaptive capacity that is often better collected using qualitative methods (Fischer et al. 2013; Wigtil et al. 2016). One way to validate such indices is to compare the post-hazard outcomes with the pre-hazard vulnerability. Therefore, in the next section, the relation between flash flood fatalities and Socio-Economic Vulnerability is investigated.

2.5.4 Flash Flood Fatality vs. Socio-Economic Vulnerability

Figure 2-10 is generated based on the data gathered by Ashley and Ashley (2008) for the period of 1959 to 2005. The fatalities were organized into three classess of high, medium, and low vulnerable. Results shows that vulnerable states are experiencing higher number of fatilities. For an instance, Arkansas has the highest fatalities and most of its counties are located in the high-vulnerable-hotspot which shows the lack of infrastructure and higher density populations in the critical counties. Nevada is among the high-

vulnerable-hotspot specifically for flash flood severity. This might be due to a better infrastructure and lower population densities in mountainous communities.

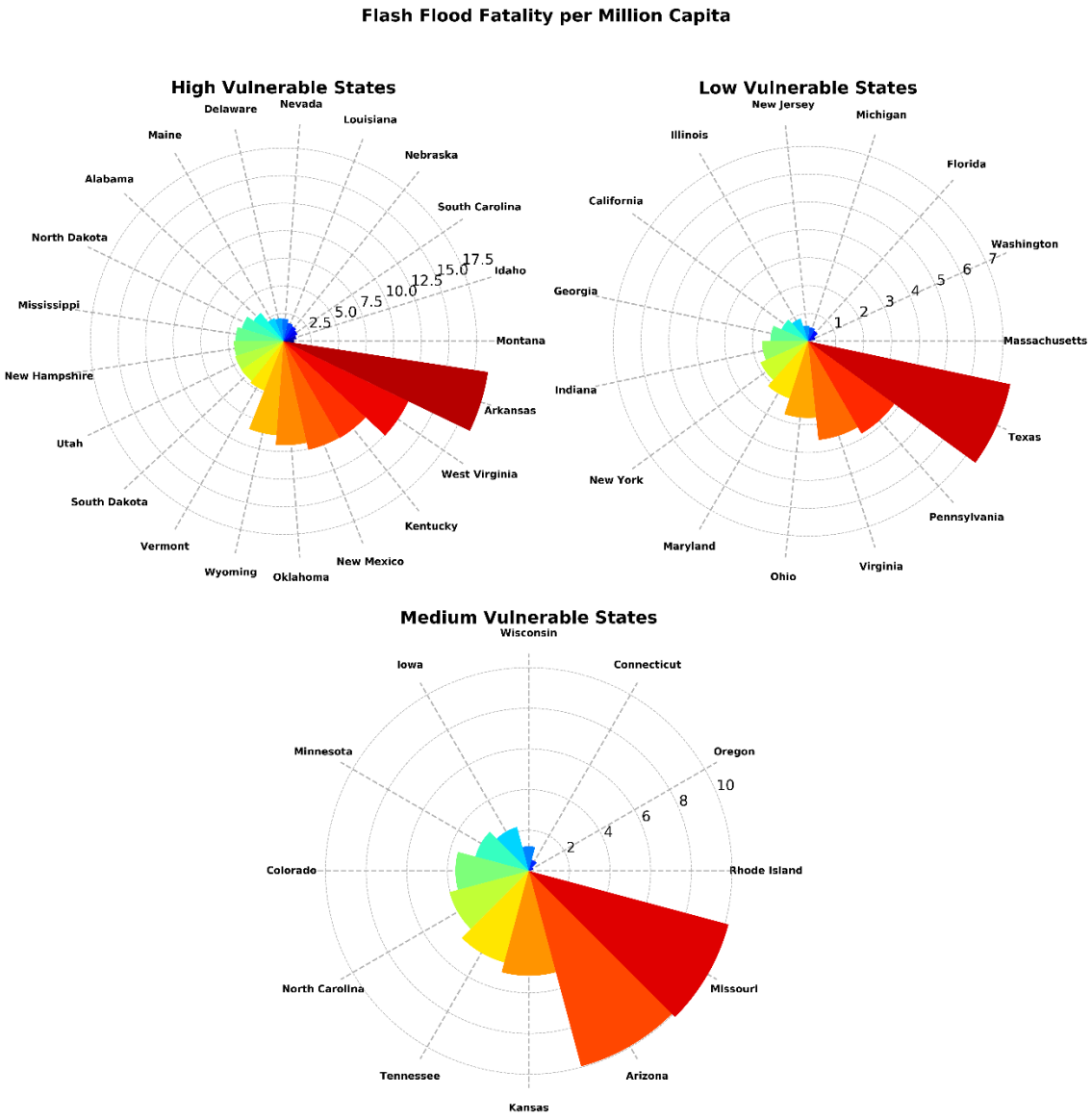

Figure 2-10. Flash Flood Fatalities in comparison to Socio-Economic Vulnerability.

Texas has experienced the highest fatality among the low vulnerable states. The counties in the lower part of Texas are among the high-vulnerable-hotspot regions, which

can explain the high fatality rate. This shows that these counties are densely populated areas.

2.6 Summary and Conclusion

This chapter presented a comprehensive assessment of socio-economic vulnerability and its interaction with flash flood characteristics over the CONUS. Vulnerability was assessed on county and state levels by employing 36 social and economic variables. Probabilistic Principal Component Analysis (PPCA) was utilized to quantify socio-economic vulnerability index (SEVI). The flash flood characteristics including frequency, magnitude, duration, and severity were considered for evaluation. Spatial distribution of flash floods were assessed using hotspot analysis. The intersection of SEVI and flash flood characteristics were mapped utilizing cross-tabulation. At last, flash flood fatalities were used to validate the calculated index. The main findings of the study are as follows:

- The spatial pattern of the SEVI correlate well with the previously reported socio-economic vulnerability. Some states such as California and Texas show low vulnerability. Whereas the majority of their counties are located in the high vulnerable regions.
- The southwest region shows severe flash flooding with high magnitude, whereas the Northern Great Plains experience lower duration and frequency.
- Critical counties (high-vulnerable-hotspot) are mostly located in the southern parts of the U.S. The majority of counties in the Northern Great

Plains are in the non-critical status. States with higher SEVI and critical counties exprience higher rates of fatalitites, such as Arkansas and Texas.

3 Assessment of the Influence of Hydroclimatic Variables on Flash Flood

3.1 Background

Flash flood is among the most hazardous natural disasters, and it can cause severe damages to the environment and human life. Flash floods are primarily caused by excessive rainfall, and due to their rapid onset (within six hours of rainfall), they can leave very limited opportunity for effective response. Climate change and global warming are expected to increase the magnitude of rainfall in some part of the world specifically in the United States, which may results in more severe flash flooding (Borga 2002; Sangati et al. 2009; Marchi et al. 2010; Borga et al. 2011; Hapuarachchi et al. 2011; Hong et al. 2013).

Compound interaction of the hydrologic and meteorological hazards is influencing the flash flooding. So far, researchers have been focusing on the development of flash flood warning systems using hydrologic simulations (Hapuarachchi et al. 2011; Hardy et al. 2016; Braud et al. 2018). However, there has been less attention toward the influencing variables that trigger flash flood. Combinations of different hydroclimatic variable such as rainfall, runoff, landcover, humidity, and soil moisture can exacerbate the occurrence of flash flood. Identifying these combinations may ease the challenges behind flash flood warning system and create a location-based perdition system.

In this chapter, a data-analytic approach is developed to assess the interaction of flash flood characteristics and the hydroclimatic variables over the CONUS. The purpose of this chapter is to identify the influencing variables on flash flood magnitude and duration at different regions across the CONUS. For this purpose, rainfall amount and rate, soil

moisture content, vegetation, precipitable water, and runoff are chosen as the influencing variables. More details about the data and methodology is provided in the next sections and the results are followed afterwards.

3.2 D-Vine Copula Based Quantile Regression Model

Regression has been implemented to quantify the relationship between dependent variable (outcome) and independent variables (covariates). Quantile regression, which is simply the prediction of conditional quantiles, has received increasing attention in recent years and it has been applied to diverse subjects such as investment, finance, engineering, and medicine. Linear quantile regression, which was introduced by Koenker and Bassett (2001), is the most common and frequently used method. Bernard and Czado (2015) proposed that the linear quantile regression is imposing too restrictive assumptions on the shape of the regression quantile. Furthermore, they discussed that the model suffers from issues like quantile crossing and the drawback of linear models such as multicollinearity, selection of significant covariates and the inclusion of interactions and transformed variables.

To overcome the pitfalls of linear quantile regression, Kraus and Czado (2017) introduced a new framework that does not make any precise assumption about the shape of the conditional quantiles. The dependence relationship between response and covariates is modeled flexibly using a parametric D-vine copula, a subclass of regular vine copulas introduced by Brechmann and Schepsmeier (2013). The main advantage of using D-vine copulas is the flexibility of separating the marginal and dependence modeling (Madadgar and Moradkhani 2013, 2014a, b; Madadgar et al. 2014; Rana et al. 2017). The proposed

algorithm has the capability of sequentially adding the covariates to the regression model with the objective of maximizing a conditional likelihood, i.e. the likelihood of the predictive response model given the covariates. On the other hand, an automatic variable selection is incorporated, meaning that the algorithm will stop adding covariates to the model as soon as none of the remaining covariates are able to significantly increase the model's conditional likelihood. This results in parsimonious and flexible models whose conditional quantiles may strongly deviate from linearity. Due to the model construction, quantile crossings do not occur. Thus, the resulting D-vine quantile regression is able to overcome all of the shortcomings of classical linear quantile regression methods that were mentioned above, and therefore adds a new (and, as it will be shown, competitive) approach to the existing research on quantile regression (Kraus and Czado 2017).

This study proposes a new framework to detect the most influential hydroclimatic variables to describe the flash flood magnitude and duration by employing the D-vine quantile regression algorithm.

3.3 Climate Forecast System

The National Centers for Environmental Prediction (NCEP) Climate Forecast System (CFS) (Saha et al. 2010, 2014) has been initiated in 2010 with the purpose of generating seasonal forecasts. The CFS project involved two models: 1) Climate Forecast System Reanalysis (CFSR), which is a fully coupled atmospheric-ocean-sea ice-land reanalysis model for 1979-2010, and it is used for creating initial conditions for Climate Forecast System version 2 (CFSv2) which retrospective forecast with the new CFS and it provides real time sub-seasonal and seasonal predictions. In this study, the hydroclimatic

variables are obtained from CFSR and CFSv2 for the period of 1979-2018 with spatial resolution of 0.5° over CONUS. Detailed description of the model products are summarized in Table 3-1.

The variables in Table 3-1 are used in the hydroclimatic evaluation of the flash flood. For further assessment of the relation between the environmental variables and flash flooding, the soil and vegetation type were also obtained from the CFS database. Detailed description of these two variables are provided in the Table 3-2 and Table 3-3.

Since climate change is among the main factors of increase in natural hazard events, the stationarity in precipitation data is evaluated to ensure the consideration of climate change in this study (Teegavarapu 2019). For this purpose, the augmented Dickey–Fuller test (ADF) introduced by Dickey (1979) is employed. ADF tests the null hypothesis that non-stationarity is present in the time series. Figure 3-1 represents the test results for seasonal precipitation over the grid-cells that are considered for this study. The majority of the United States shows non-stationarity in the CFS data, whereas the central region and South west are indicating stationarity.

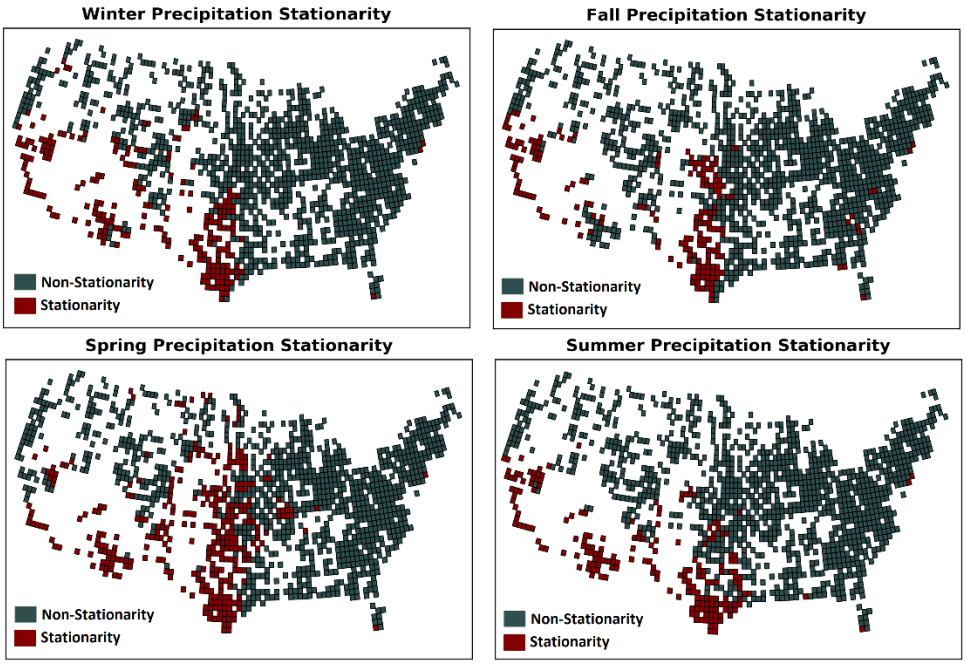

Figure 3-1. Stationarity analysis of seasonal precipitation from CFS.

3.4 Flash Flood Characteristics

The USGS automated measurements for the Unified Flash Flood Database developed in HyDROS are used to extract the flash flood magnitude and duration for the period of 1980-2016 for 2,751 USGS gauges over the CONUS. Detailed description of this dataset is described in section 3.2. Since the flash flood data are archived in station-based records and this study aims to build the model according to the CFS model and flash flood characteristics, the stations were converted to distributed model with 0.5° spatial resolution. Therefore, the grids in the gridded flash flood record are consistent with the CFS resolution. Due to the density of stations in some regions, some of the grids contained more than one station. Figure 3-2 shows the gridded flash flood data developed in this study.

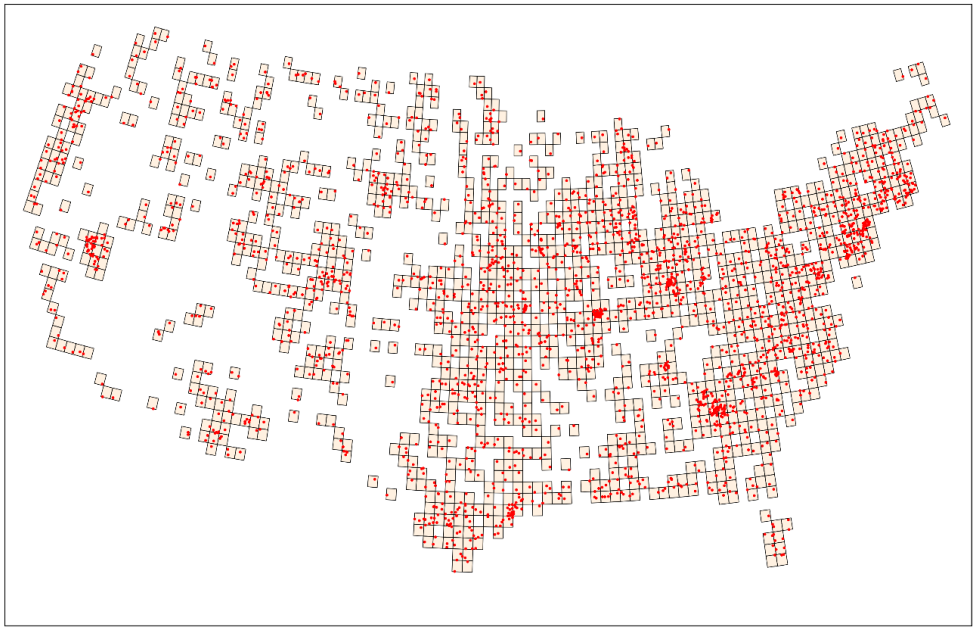

Figure 3-2. Gridded USGS Stations.

Table 3-1. The hydroclimatic variables obtained from the CFSv2 database

Variables	Description
Precipitation rate (kg m-2 s-1)	6-hour Average (initial+0 to initial+6)
	0.312° × ~0.312° from 0E to 359.688E and 89.761N to 89.761S (1152 × 576 Longitude/Gaussian Latitude)
Soil Moisture Content (kg m-2)	6-hour Forecast
	Layer between two 'Depth below land surface': Bottom=2 m, Top=0 m
	0.312° × ~0.312° from 0E to 359.688E and 89.761N to 89.761S (1152 × 576 Longitude/Gaussian Latitude)
Storm Surface runoff (kg m-2)	6-hour Accumulation (initial+0 to initial+6)
	0.312° x ~0.312° from 0E to 359.688E and 89.761N to 89.761S (1152 x 576 Longitude/Gaussian Latitude)
Vegetation (%)	6-hour Forecast
	0.312° x ~0.312° from 0E to 359.688E and 89.761N to 89.761S (1152 x 576 Longitude/Gaussian Latitude)
Water Runoff (Kg m-2)	6-hour Accumulation (initial+0 to initial+6)
	0.312° x ~0.312° from 0E to 359.688E and 89.761N to 89.761S (1152 x 576 Longitude/Gaussian Latitude)
Precipitation amount (kg m-2)	6-hour Average (initial+0 to initial+6)
	0.312° x ~0.312° from 0E to 359.688E and 89.761N to 89.761S (1152 x 576 Longitude/Gaussian Latitude)
Precipitable water (kg m-2)	6-hour Average (initial+0 to initial+6)
	0.312° x ~0.312° from 0E to 359.688E and 89.761N to 89.761S (1152 x 576 Longitude/Gaussian Latitude)

3.5 Methodology

3.5.1 Pair Copula Construction

Copula functions were introduced by Sklar (1959) and Joe (1997) as multivariate functions that capture the dependence structure between random variables regardless of

their marginal distributions. Since the copula functions are built on the Kendall's tau correlations, they overcome the distributional assumptions made in linear models and they are practical in modeling non-linear relationships between different variables (Khajehei and Moradkhani 2017; Khajehei et al. 2018).

Mathematically, copula is defined as the joint cumulative density function (cdf) of two univariate marginal distribution functions (Joe 1997; Nelsen 1999). The multivariate copula C is defined as:

$$C: [0,1]^n \rightarrow [0,1] \quad \textbf{(3-1)}$$

In an n dimensional space, the continuous variables X_i are defined by their own cdfs and the probability density functions (pdfs) named as $F_i(x_i)$ and $f_i(x_i)$, respectively. The joint cdf of the random variables can be written as:

$$F(x_1, x_2, \dots, x_n) = C[F_1(x_1), F_2(x_2), \dots, F_n(x_n)] \quad \textbf{(3-2)}$$

And the corresponding pdf is formulated as following:

$$f(x_1, \dots, x_n) = c\big(F_1(x_1), F_2(x_2), \dots, F_n(x_n)\big)\prod_{i=1}^{n} f_i(x_i) \quad \textbf{(3-3)}$$

The copula density is described as below where $u_i = F_i(x_i)$ with $u_i \in [0,1]$:

$$c(u_1, \dots., u_n) = \frac{\boldsymbol{\partial^n} C(u_1, \dots, , u_n)}{\boldsymbol{\partial} u_1 \dots \boldsymbol{\partial} u_1} \quad \textbf{(3-4)}$$

As shown in equation 3-2, the copula functions are capable of modeling the dependence structure separately with the marginal distribution.

Copula functions are categorized into several categories and two of the most common categories that are being used in hydroclimatic research are Archimedean and

Elliptical. Elliptical copulas are capable of capturing the pair-wise correlations among the variables with any level of correlation; however, they do not have a closed form expression and the copula functions in each level must belong to the same family, where the estimated parameter in the high levels are smaller than those in the lower levels (Zhang and Singh 2014). On the other hand, the Archimedean copulas provide the closed form expression but cannot preserve all the pair-wise correlations and most importantly only a few of the bivariate copulas from this group can be extended to the higher than two dimensions space.

To overcome these shortcoming from copula families, Bedford and Cooke (2001) extend the work by Joe (1997) that introduced the vine copulas. Vine copula has three forms including Regular (R)- Vine, Canonical (C)- Vine, and D-Vine. Vine copulas were introduced to help organizing the pair-copula construction (PCC). A PCC of a three-dimensional density is described as below:

$$f(x_1, x_2, x_3) = f(x_1).f(x_2).f(x_3).c_{12}\big(F(x_1), F(x_2)\big) \\ .c_{23}\big(F(x_2), F(x_3)\big).c_{13|2}\big(F(x_1|x_2), F(x_3|x_2)\big) \quad \textbf{(3-5)}$$

In the case of five-dimensional density, there are as many as 480 different such construction as Equation (3-5), and 23,040 for the 6-dimensional case and so on. Hence, for higher dimensional distributions, there are a significant number of possible pair-copula constructions, graphical models called R-vines were introduced by Bedford and Cooke (2001) as below:

A regular vine is a sequence of d-1 linked trees where: 1) Tree T_1 is a tree on nodes 1 to d. 2) Tree T_n has d-1-j nodes and d-j edges. 3) Edges in tree T_n become nodes in tree

T_{n+1}. 4) Two nodes in tree T_{n+1} can be joined by an edge only if the corresponding edges in tree T_n share a node.

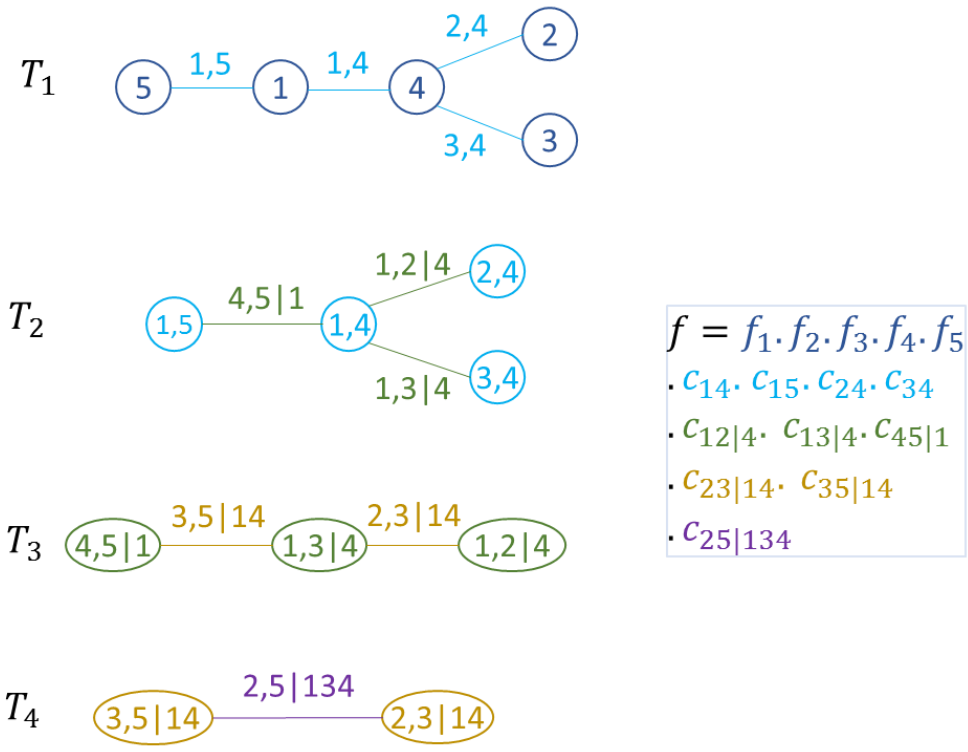

Figure 3-3. Schematic Diagram Representing Five Dimensional Copula Density with R-Vines.

The C-vine is a special case of R-vine, whereas each tree has unique node that is connected to n-j edges. C-vine is beneficial to be employed where there is a key variable that governs the interactions in the dataset.

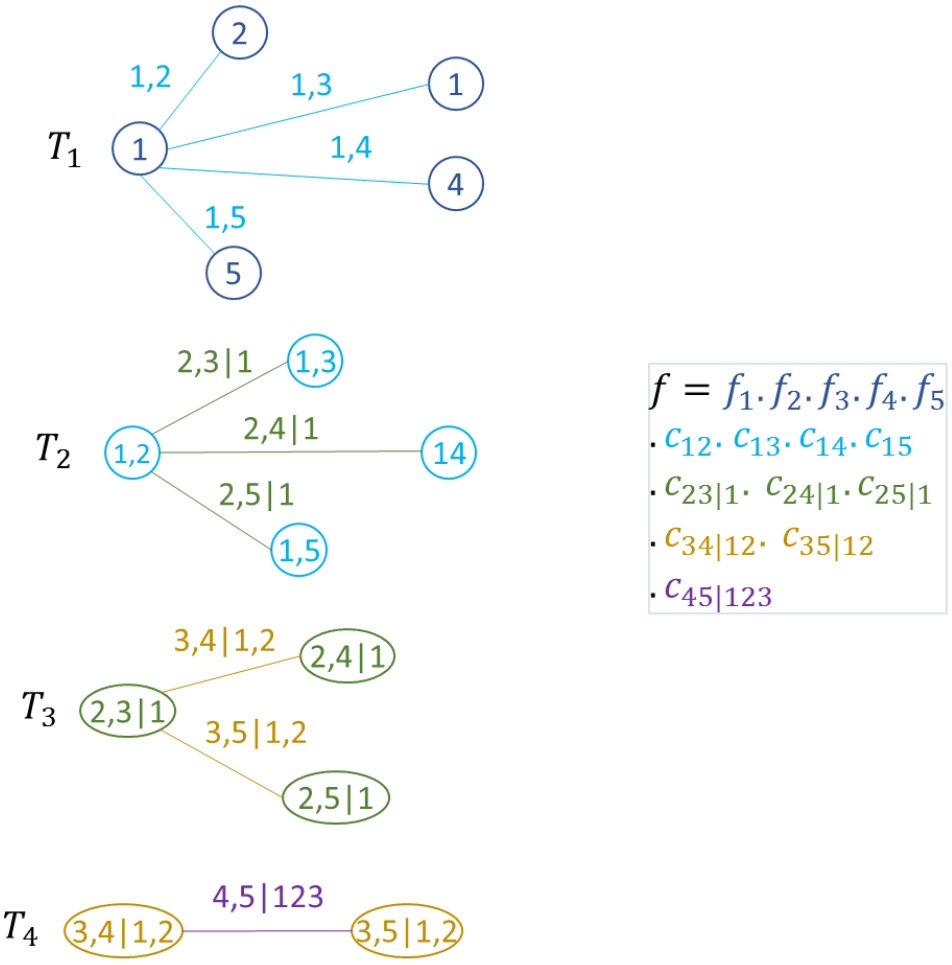

Figure 3-4. Schematic Diagram Representing Five Dimensional Copula Density with C-Vines.

The D-vine copula is a special case of R-vine, whereas no node in any tree is connected to more than two edges. The D-vine is a hierarchical model that resembles independent graphs more than the C-vine copula. The general expression for both the C-vine and D-vine structures is the same for less than three-dimensional cases.

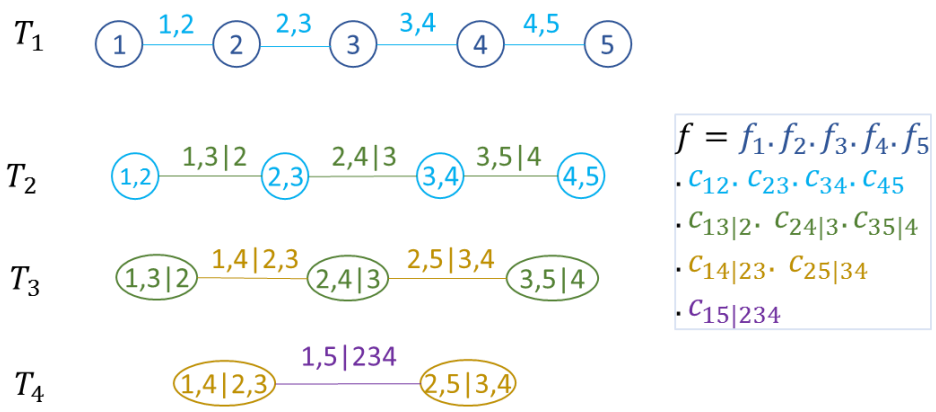

Figure 3-5. Five Dimensional Copula Density with C-Vines.

3.5.2 D-Vine Quantile Regression Model Construction

The aim of this chapter is to find the influencing variables on flash flood in order to be employed in forecasting the conditional quantile q_a of the response variable (i.e, flash flood events) at some arbitrary level $a \in (0.1)$ for the given covariates (i.e., chosen hydroclimatic variables). For this purpose, Kraus and Czado (2017) employed the inverse of conditional distribution based on D-vine copula as follows:

$$q_a(X_1, \dots, X_n) = F^{-1}_{Y|X_1,\dots,X_n}\langle a|x_1, \dots, x_n\rangle \quad \textbf{(3-6)}$$

By using the probability integral transform, where $V = F_Y(Y)$ and $V = F_n(X_n)$, the right hand side of Equation (3-6) can be formulated as:

$$F_{Y|X_1,\dots,X_n}\langle y|x_1, \dots, x_n\rangle = C_{V|U_1,\dots,U_n}\langle v|u_1, \dots, u_n\rangle \quad \textbf{(3-7)}$$

$$F^{-1}_{Y|X_1,\dots,X_n}\langle y|x_1, \dots, x_n\rangle = C^{-1}_{V|U_1,\dots,U_n}\langle v|u_1, \dots, u_n\rangle \quad \textbf{(3-8)}$$

Therefore, estimating the conditional quantile function can be obtained by estimating the marginals F_Y and F_j, j=1,…,n.

Based on the above theory, this study adopted the D-vine copula regression model to select the most influential hydroclimatic variables on flash flood characteristics. The steps taken to build the model for this purpose are described in the following.

3.5.2.1 Marginal Distribution Estimation

Assume an n-dimensional dataset (Y, X_1,..,X_n) where Y represents the flash flood characteristics and X denotes the hydroclimatic variables. All the variables need to be transferred to pseudo-copula data (V, U_1,…,U_n) employing the variables' marginal distributions. As shown in Equation (3-8), the inverse of the estimated marginal distributions for the quantile regression model in required; therefore, the parametric distributions are used for the marginal distribution of flash flood characteristics and hydroclimatic variables. For this purpose, eight commonly used distributions including Gaussian, Lognormal, Generalized Extreme Value (GEV), Generalized Pareto (GP), Gumbel, Gamma, and Exponential are employed. Bayesian Information Criterion (BIC) (Aho et al. 2014) and Kolmogorov-Smirnov (K-S) (Stephens 1974) tests are utilized to determine the suitable distribution for each case.

The Kolmogorv-Smirnov (K-S) test is among the nonparametric tests that is used to choose the reference distribution that is suitable for a given dataset (Stephens 1974).

$$D = Max\{|F(x) - G(x)|\} \quad \textbf{(3-9)}$$

where F(x) and G(x) are the empirical and reference CDFs, respectively. The null hypothesis in K-S test is that the reference distribution is able to describe the given datasets (observation and forecast). With the significance level of α=0.05, if the calculated p-value

is larger than α, the null hypothesis is not rejected and that specific distribution will be chosen among the approved parametric distributions.

Then, Bayesian Information Criterion (BIC) (Aho et al. 2014) is used to select the best distribution from the approved distributions by the K-S test. The reference distribution with the lowest BIC value is the desirable distribution.

$$BIC = -2 \times ln(L) + k \times ln(n) \quad \textbf{(3-10)}$$

where L is the maximized value of the likelihood function, k is the number of parameters in the model, and n is the number of data points in the dataset.

3.5.2.2 Step-wise D-Vine Copula Estimation

Equation (3-8) is utilized to calculate the conditional quantile. Therefore, a D-vine with ordering of $l = (l_1, \dots, l_n)$ is fitted to the pseudo copula data (V, $Ul_1,\dots,Ul_n$). The ordering can be chosen arbitrarily. However, the explanatory power of the final model is affected by the particular ordering, and the aim is to build a quantile regression model with the most explanatory power. As a result, Kraus and Czado (2017) proposed an automated algorithm that constructs the D-vine sequentially choosing the most influential covariates. The algorithm starts with zero covariates, and in each step, a new covariate is added to the model that improves the model's fit the most. The model fit can be evaluated with two techniques, the corrected AIC (Akaike Information Criterion) conditional log-likelihood (cll_{AIC}) and the corrected BIC conditional log-likelihood:

$$cll_{AIC}(l, F, \theta, V, U) = 2cll(l, F, \theta, V, U) + 2|\theta| \quad \textbf{(3-11)}$$

$$cll_{BIC}(l, F, \theta, V, U) = -2cll(l, F, \theta, V, U) + log(n)\,|\theta| \quad \textbf{(3-12)}$$

Whereas the conditional likelihood function is defined as below:

$$cll(l, F, \theta, V, U) = \sum_{i=1}^{d} ln\, c_{V|U}\left(v^{(i)}\middle|u^{(i)}: l, F, \theta\right) \quad \textbf{(3-13)}$$

In which the conditional $c_{V|U}$ can be formulated as following (Killiches et al. 2018):

$$c_{V|U}\left(v^{(i)}\middle|u^{(i)}: l, F, \theta\right) = c_{VU_{L_1}}\left(v^{(i)}\middle|u_{l_1}^{(i)}: F_{VU_{l_1}}, \theta_{VU_{l_1}}\right) \times$$

$$\prod_{j=2}^{n} c_{VU_{l_j}}; U_{l_1}, \dots, U_{l_{j-1}}\left(C_{V|U_{l_1},\dots,U_{l_{j-1}}}\left(v^{(i)}\middle|u_{l_1}^{(i)}, \dots, u_{l_{j-1}}^{(i)}\right),\right. \quad \textbf{(3-14)}$$

$$\left. C_{U_{l_j}|U_{l_1},\dots,U_{l_{j-1}}}\left(U_{l_j}{}^{(i)}\middle|u_{l_1}^{(i)}, \dots, u_{l_{j-1}}^{(i)}\right); \mathrm{F}_{\mathrm{VU_{l_j};U_{l_1},\dots,U_{l_{j-1}}}}, \theta_{\mathrm{VU_{l_j};U_{l_1},\dots,U_{l_{j-1}}}}\right.$$

Where F and Θ represent the estimated family and parameters of pair-copula c, respectively. This study uses the $\mathrm{cll_{AIC}}$ as the measurement for the model's fit.

3.6 Results and Discussion

This study implements the D-Vine copula technique to find the influential hydroclimatic variables on the two flash flood characteristics (i.e., duration and magnitude). Results from this study are summarized in two section. The first section focuses on flash flood duration, and the flash flood magnitude outputs are presented in the second section. Since vegetation and soil type can affect the flash flood occurrence and they are not time varying variable, the third section is dedicated to evaluate the interaction between flash flood characteristics and these two physical properties.

3.6.1 Influential Variables on Flash Flood Duration

Before evaluation of D-Vine model results, the relationship between flash flood duration and hydroclimatic variables is assessed. Figure 3-6 shows the correlation between each variable and flash duration at each station. The purple and red colors show significant correlation and the low correlation is shown in green. Comparing the results for the hydroclimatic variables indicates there are number of stations with positive correlation and there are stations that have negative values. For instance, the precipitation amount and precipitation rate have positive correlation with flash flood duration in most stations in the Pacific region, whereas; the vegetation has a negative correlation in that region. Overall, soil moisture demonstrates significant positive correlation in the majority of stations; whereas, vegetation indicates negative correlation in most of the regions.

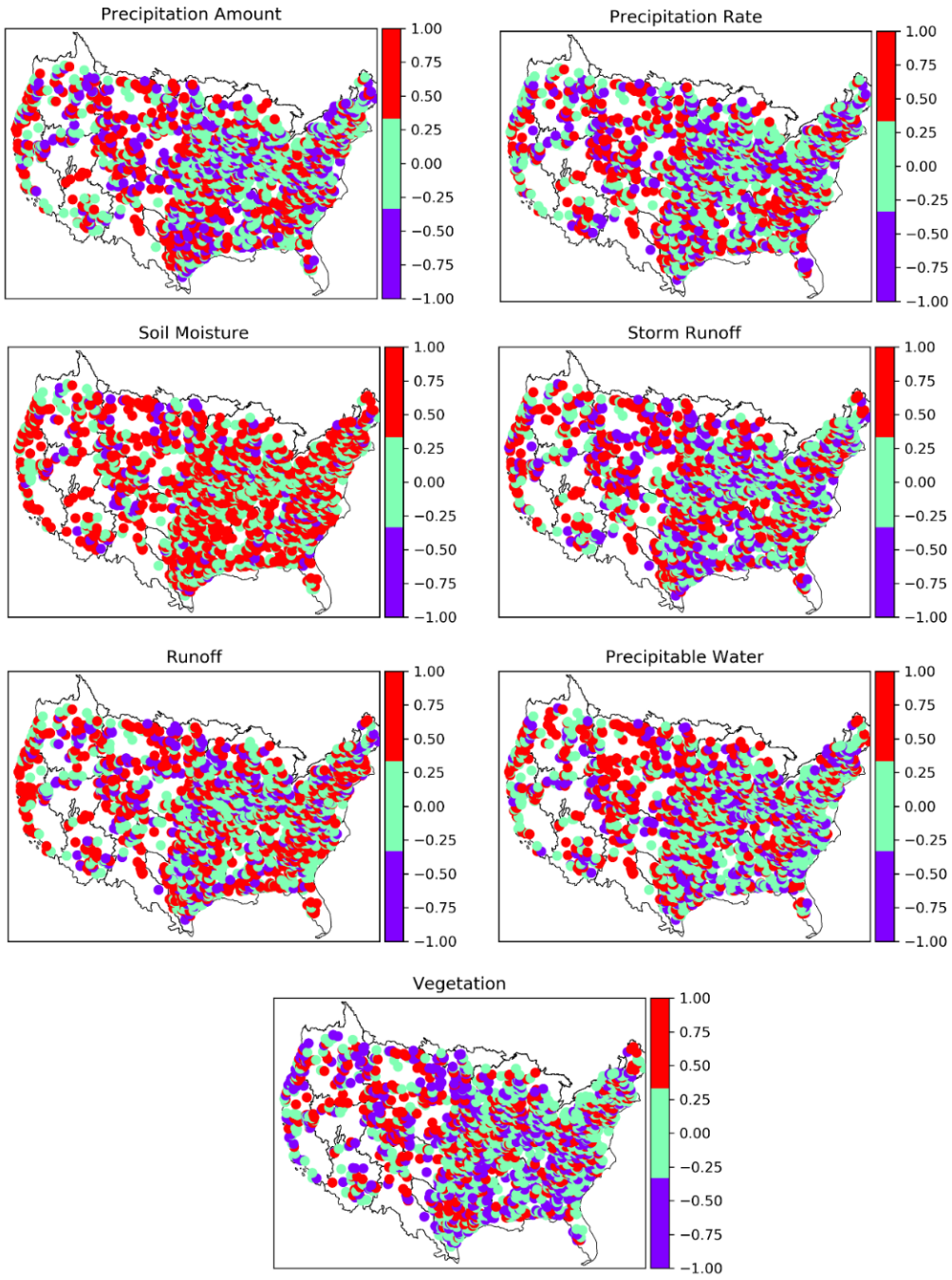

Figure 3-6. Kendall's Tau-A correlation between flash flood duration and the chosen hydroclimatic variables, precipitation amount and rate, solid moisture, runoff, storm runoff, precipitable water, and vegetation.

However, relying on the results from correlation statistics is not sufficient to draw conclusion on the relationship between two variables. In this study, since the flash events are analyzed, the time series are not continuous for the study period, therefore this causes bias in the calculated correlation. Furthermore, the number of events are not large enough is some regions to calculate a significantly meaningful correlation. Lastly, correlation does not mean causation. Therefore, there is a need to investigate the interaction between these variables with a more advanced approach.

The results from the D-Vine copula model are summarized in Figure 3-7 through Figure 3-21. Since this model is simulated over 2,710 stations, showing the model statistics such as marginal distributions and the chosen copula functions is not feasible. In other words, the copula models are tested separately for each station and the best model is chosen for each case according to the objective functions. Therefore, this section focuses on the model outputs and the spatial distribution of influential variables.

Figure 3-7 represents the stations that each particular variable is chosen among the influential variables. Precipitation amount shows a disperse spatial distribution comparing to the rest of variable and it was chosen in 1,112 stations. Precipitation rate is among the influential variables in 411 stations. This implies that precipitation amount has more effect on the flash flood comparing to the precipitation rate. Storm runoff and runoff have approximately the same spatial distribution with 407 and 467 chosen stations, respectively. Soil Moisture is the second highly chosen variable after precipitation amount with 1,056 stations. Soil moisture initial condition and rainfall are among the most important influences in flash flooding (Norbiato et al. 2008; Sangati et al. 2009), which is also the

case in this study. Lastly, vegetation and precipitable water are affecting the flash flood duration in 595 and 585 number of stations, respectively.

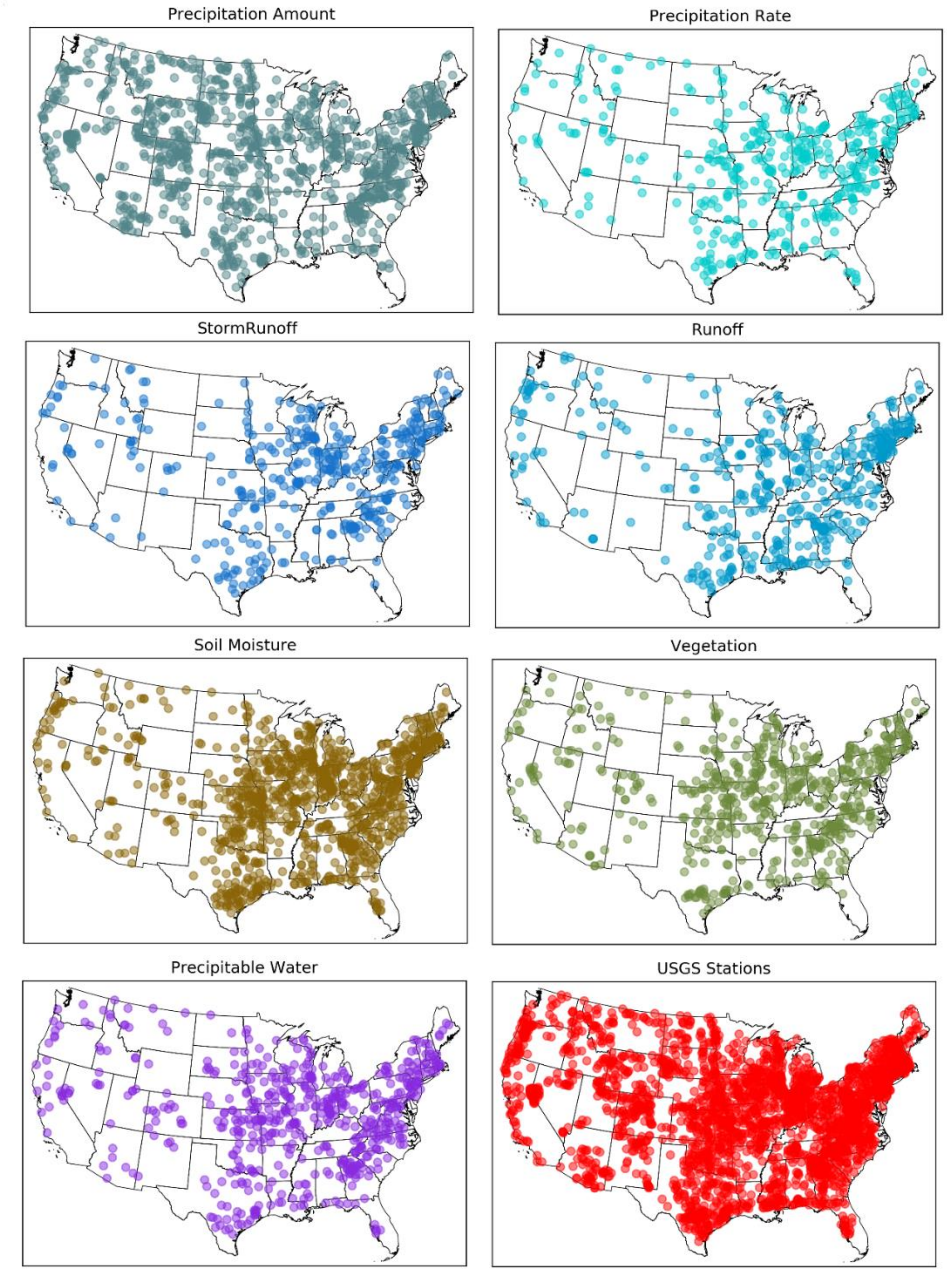

Figure 3-7. The hydroclimatic variables influencing the flash flood duration at different USGS station across the CONUS.

To study the characteristics of the D-Vine Copula model, assessing the number of predictors (i.e., covariates) that is chosen at each station is necessary. Figure 3-8 demonstrates the results for the number of predictors at each station. The models with one predictor covers the majority of the regions in CONUS with 1,377 stations. The number of predictors has a strictly decreasing pattern with 933 stations for two predictors, 381 stations for three predictors, and 60 stations indicating suitability of four and five predictors to construct the copula model. The model does not go beyond five predictors, even though there are seven variables available.

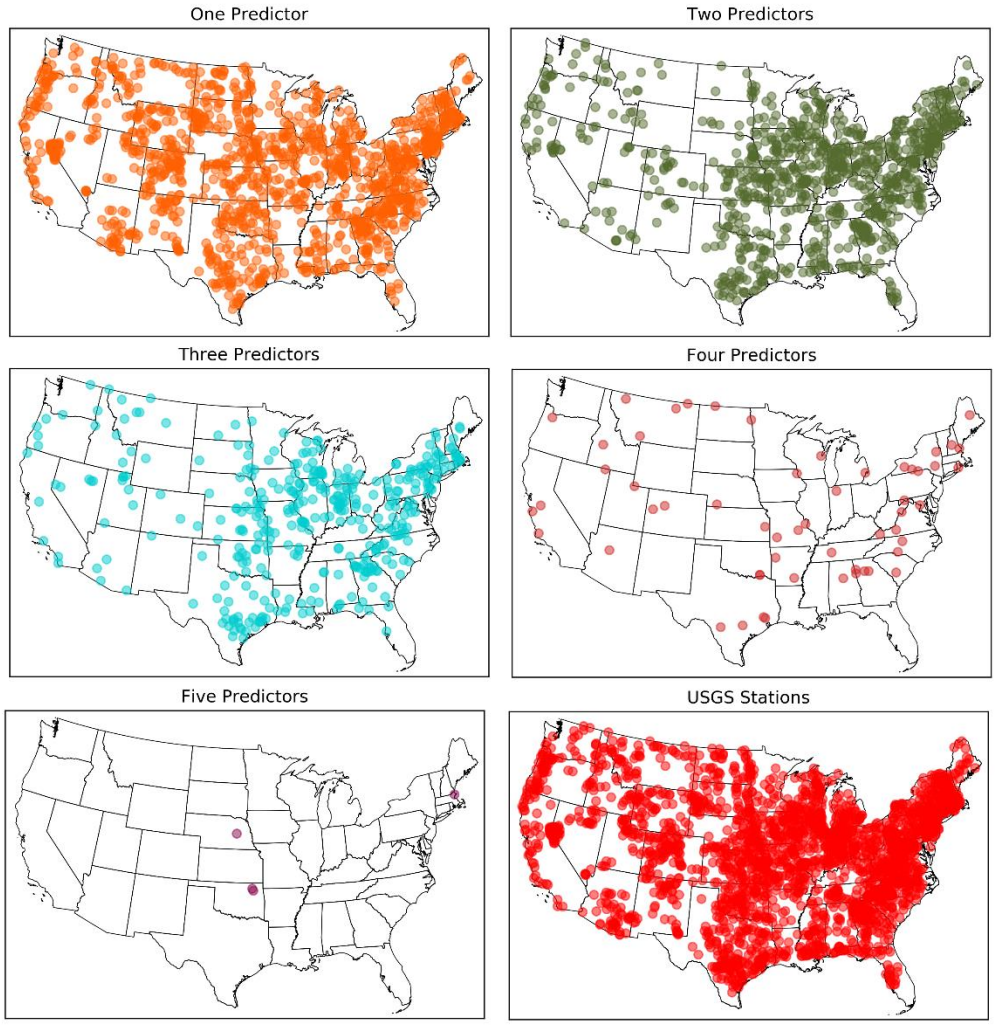

Figure 3-8. Number of predictors (i.e., covariates) chosen in eaah USGS station to construct the D-vine copula model for flash flood duration.

For detailed evaluation of the the behaviour of the predictors and D-Vine copula model, the stations are categorized according to 18 major river basins defined by the U.S. Geological Survey (Seaber et al. 1987). Figure 3-9 and 3-10 demonstrate the percentage of stations located in a certain region with a specific number of covariates (i.e., predictors) in the D-Vine copula model.

Figure 3-9 summarized the results for the regions 1-9 and the rest are shown in Figure 3-10. As Figure 3-8 is showing, majority of the regions have more that 30 percent stations with single predictor model. Among different regions, region 13 has the highest percentage of single predictor model with approximately 30 percent, followed by regions 14 and 15 with almost 70 percent. Two predictor models are covering fewer stations with the highest rate of 50 percent in region six. In three regions (i.e. regions 6, 7, and 8), the number of staions with two predictors are higher than one predictor stations. Between 5 to 20 percent of the stations are dominated by three-predictor model. Regions 9, 13, and 14 have the smallest rate of three predictors, whereas regions 7 and 12 indicate more that 15 percent of such. However, none of the regions is dominated by higher than 2 predictors. In fact, regions 5 and 13 are goverened by three or fewer predictors. In region 9, the fourpredictor model is covering more stations comparing to three predictors.

Focusing on the spatial pattern of the number of predictors, approximately most of the nearby regions follow a similar pattern. For instance, the ratios for the regions in eastern U.S. are similar and they are significantly different from the central and western regions. Western regions are highly dominated by one predictor, whereas the difference between one predictor and two-predictor models are not as considerable in the eastern regions.

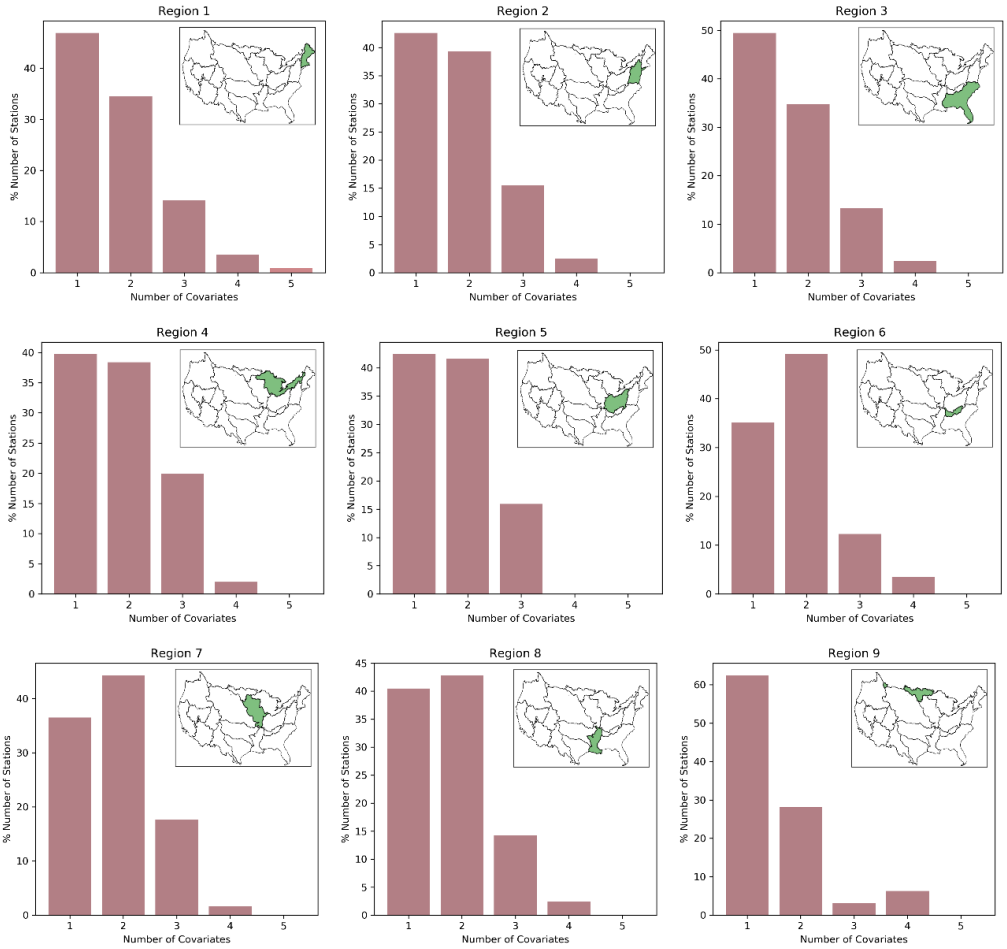

Figure 3-9. Percentage of stations dominated by a specific number of predictors (covariates) in USGS water management regions 1 to 9.

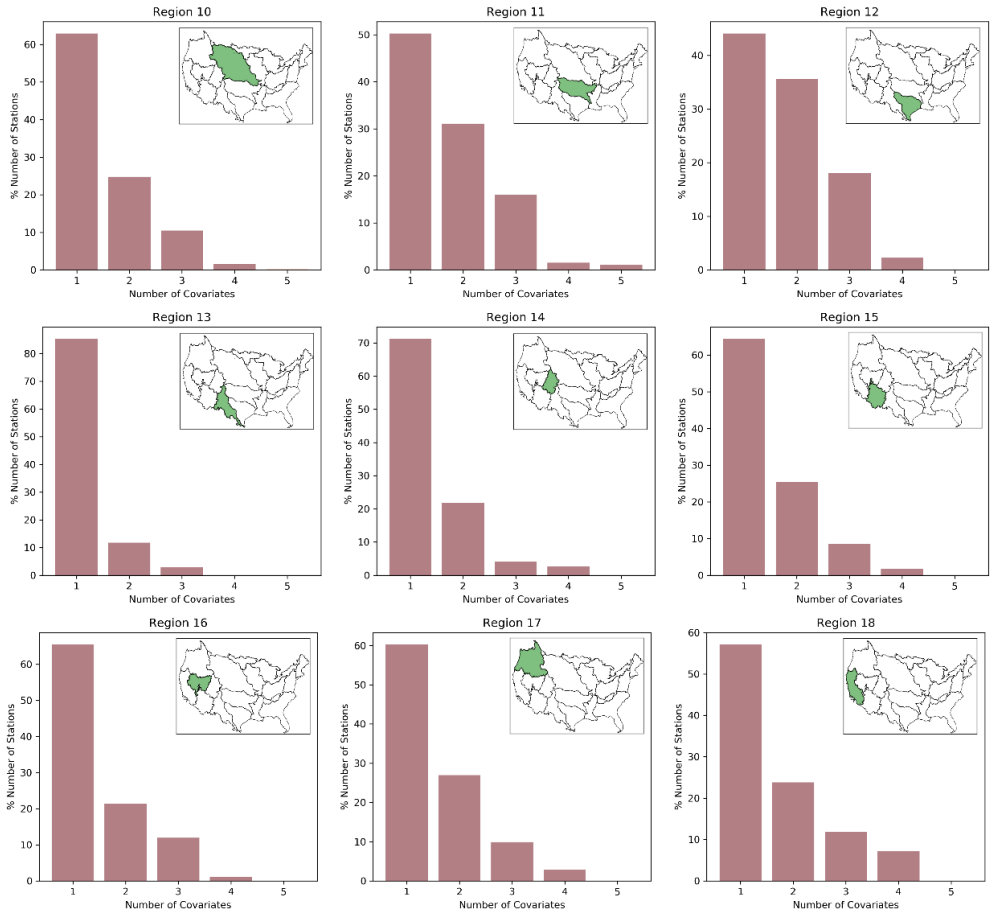

Figure 3-10. Percentage of stations dominated by specific number of predictors (covariates) in USGS water management regions 10 to 18.

Figure 3-11 and Figure 3-12 represent the percentage of stations with the specified influential variables. Precipitation amount is dominating eight regions including regions 9, 10, 13, 14, 15, 16, 17, and 18. Regions 13 and 14 are dominated by precipitation amount in 80 percent of their stations. Ogden et al (2000) indicated the high effect of precipitation amount on the flash flood in one of the basins in Colorado, which coroborate the impact of precipitation on flash flood in region 14. Furthormore, Maddox et al. (1980) investigated the meorological charateristics of flash flood in Western United States and showed the higher influence of rainfall in the eastern regions (i.e., region 13). In majority of the eastern

regions, storm runoff dominates the model with more than 30 percent of the stations. Soil moiture, runoff, precipitable water, and vegetation are dominating between 10 to 30 percent of stations in the eastern regions.

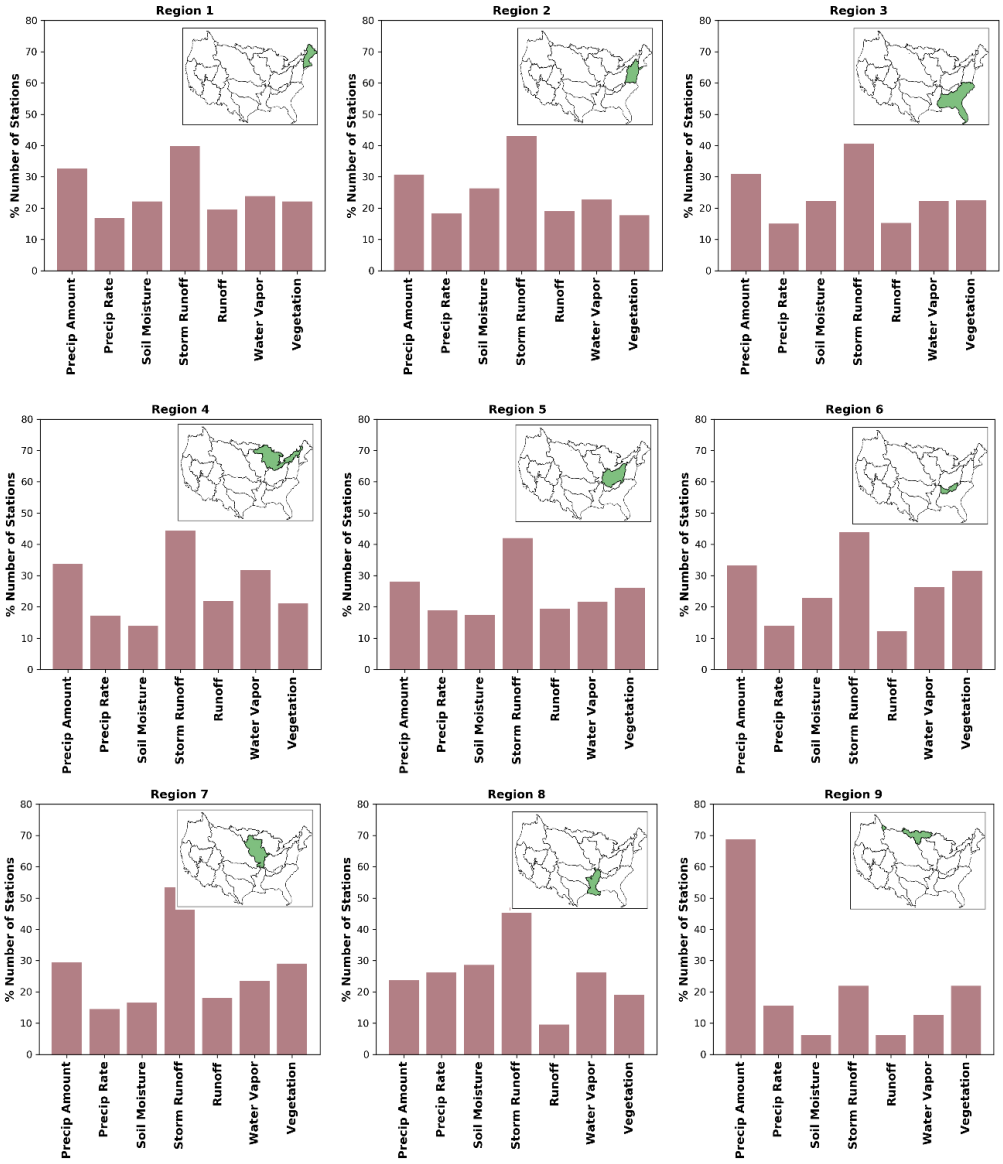

Figure 3-11. Percentage of stations with specific influencing hydroclimatic variable in USGS water management regions 1 to 9.

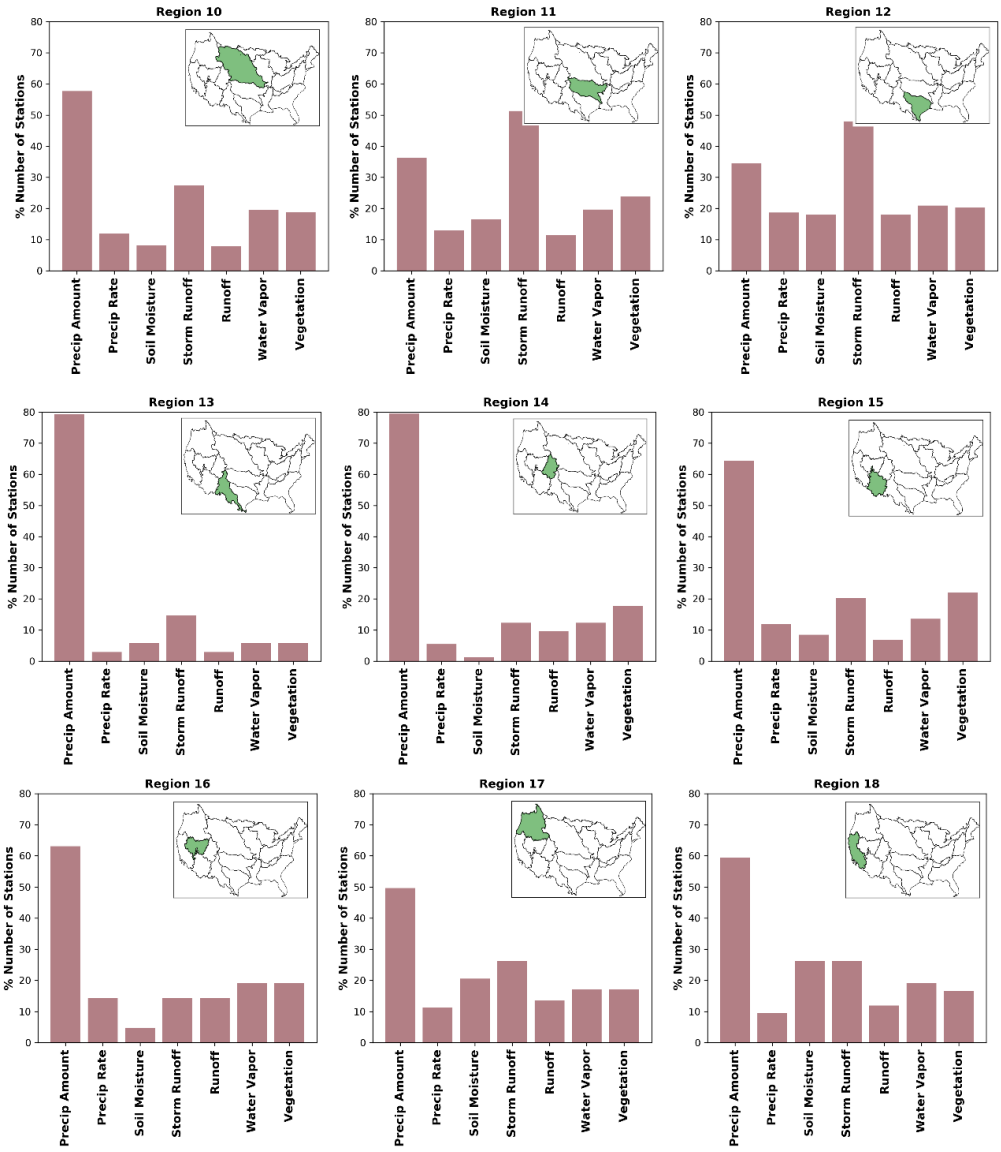

Figure 3-12. Percentage of stations with specific influencing hydroclimatic variable in USGS water management regions 10 to 18.

3.6.2 Influential Variable on Flash Flood Magnitude

Figure 3-13 shows the correlation between each variable and flash magnitude in each station. The purple and red colors show significant positive and negative correlations, respectively, and insignificant correlation is shown in green. Comparing the results for the hydroclimatic variables, there are number of stations with positive correlation and there are stations that has negative correlation with the flash flood magnitude. For instance, the precipitation amount and precipitation rate have positive correlation with the flash flood magnitude in the majority of stations in the Pacific region. Whereas, vegetation has a negative effect in that region. Overall, soil moisture indicates significant positive correlation in majority of stations. However, vegetation has a negative effect on flash flood duration in most regions.

As it is described in section 3.6.1, relying on the results of correlation statistics is not sufficient to draw conclusion on the relationship between two variables. Similar to what was carried out for flash flood duration, the D-Vine copula model is employed again to find the influencing variable on flash flood magnitude.

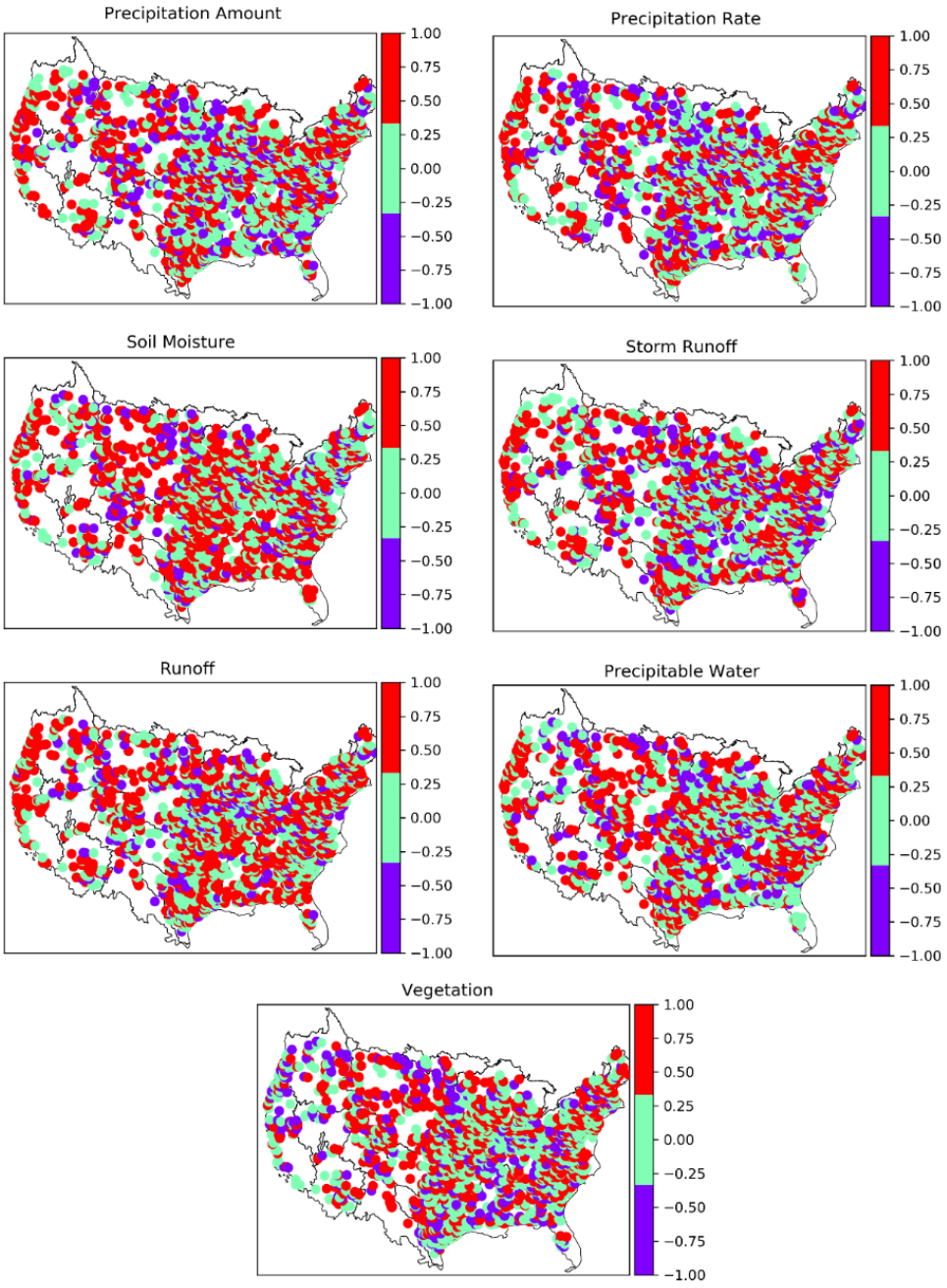

Figure 3-13. Kendall's Tau-A correlation between the flash flood magnitude and the chosen hydroclimatic variable, precipitation amount and rate, solid moisture, runoff, storm runoff, precipitable water, and vegetation.

Figure 3-14 presents the stations with a specific variable chosen among the influential variables on flash flood magnitude. Precipitation amount has a disperse spatial distribution comparing to the rest of variable and it was chosen among 1,133 stations. Precipitation rate is among the influential variables in 420 stations. This shows that precipitation amount has more effect on flash flood comparing to the precipitation rate. Storm runoff and runoff have approximately the same spatial distribution with 580 and 462 chosen stations, respectively. Results show that runoff has more influence on flash flood magnitude comparing to the duration. Soil Moisture is the second most chosen variable after precipitation amount with 854 stations, which is less than the rate for duration. Vegetation and precipitable water are affecting the flash flood duration in 593 and 618 stations, respectively. These results show that flash flood magnitude is more sensitive to the factors that affect the runoff amount, like rainfall and precipitable water. However, soil moisture can control the time that water infiltrates the soil, and therefore, it controls flash flood duration in more regions comparing to magnitude.

Figure 3-15 demonstrates the results for the number of predictors at each station for the flash flood magnitude. The models with one predictor covers the majority of the regions in CONUS with 1,370 stations. The number of predictors has a strictly decreasing pattern with 917 stations with two predictors, 415 stations with three predictors, and 54 stations experiencing four and five predictors. The model characteristics based on the number of predictors are similar to the ones from the flash flood duration.

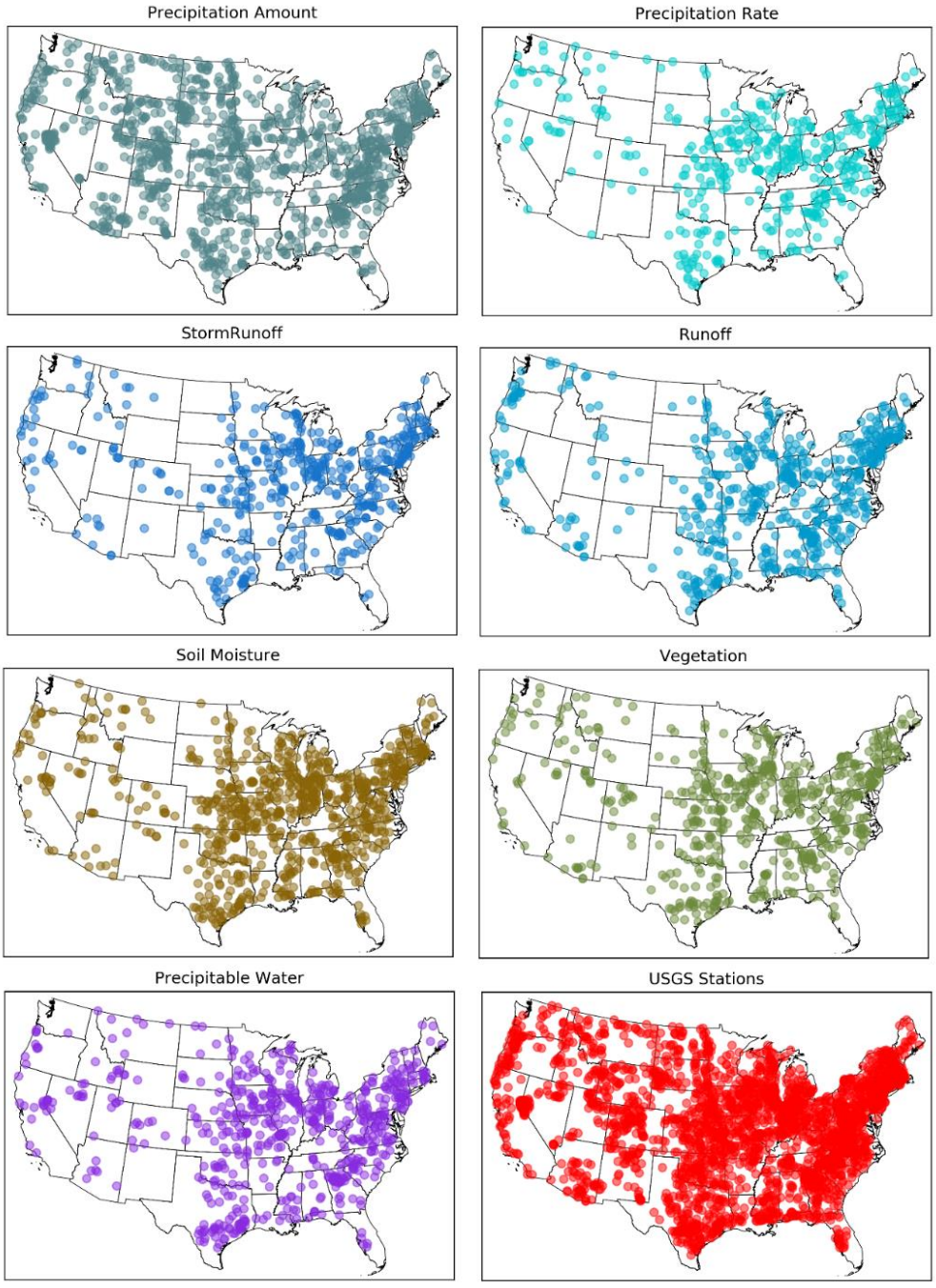

Figure 3-14. The hydroclimatic variables influencing the flash flood magnitude at different USGS station across the CONUS.

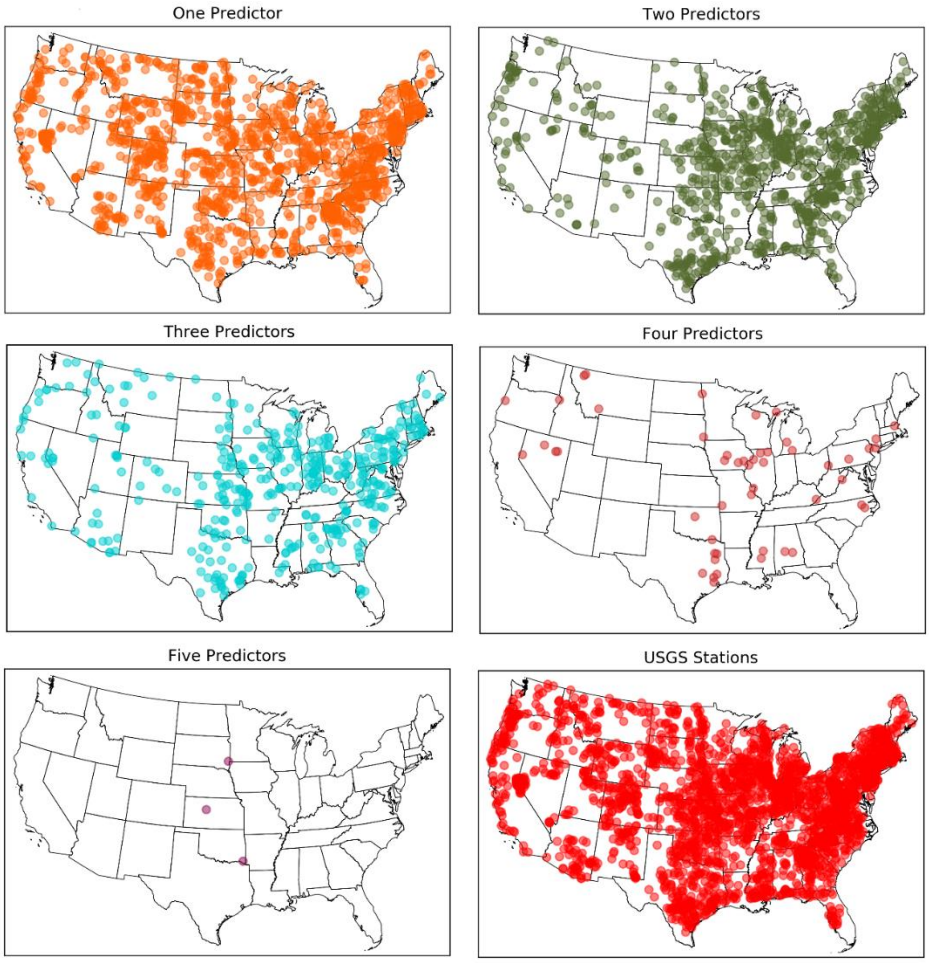

Figure 3-15. Number of predictors (i.e., covariates) chosen at the USGS stations for flash flood magnitude.

Figure 3-16 and 3-17 present the regional distribution of D-Vine Copula model based on the number of predictors in the model. One-predictor models dominate most of the regions except for region 6, 7, and 12. More than 80 percent of the stations in regions 13 and 14 are governed by one predictor model. Region 6, 7, and 12 are dominated by two predictors with approximately 45 percent of the stations. The percentage of stations in region 5 shows equal one and two predictor models. Three predictors models are shown in less than 20 percent of stations in each region. In region 15, three predictors models owns more than 10 percent of the stations comparing to two predictors.

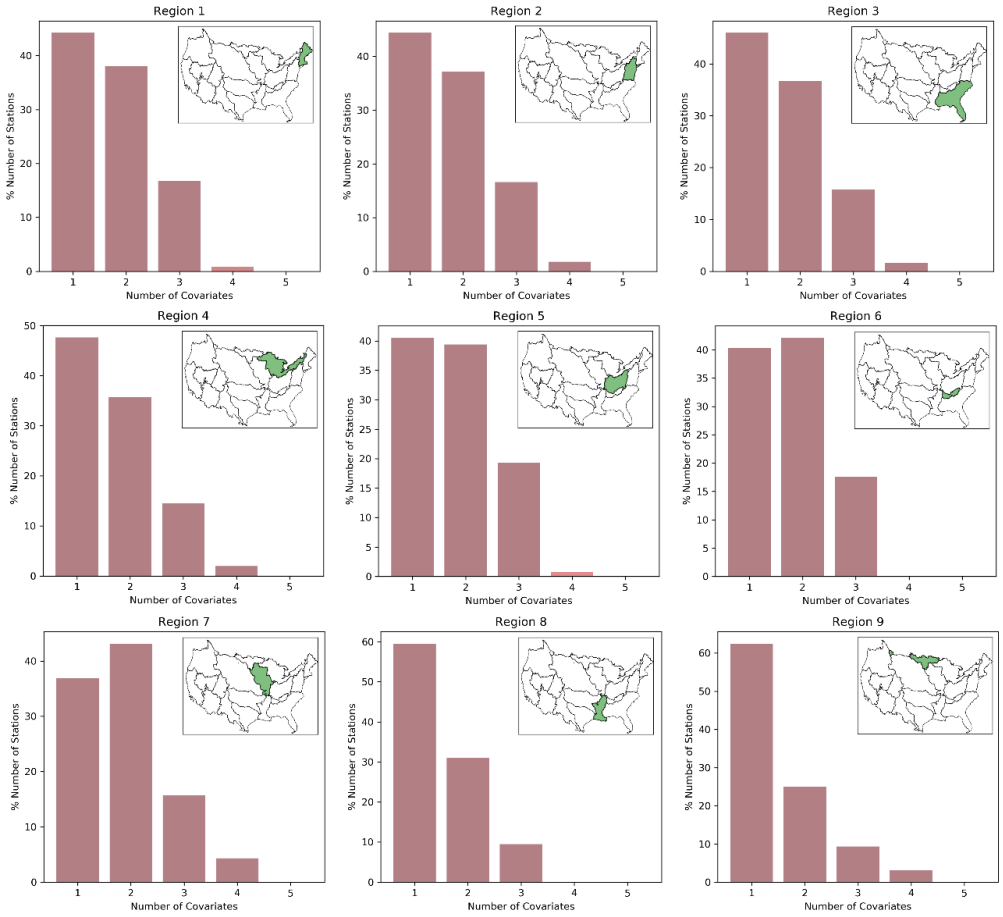

Figure 3-16. Percent of stations dominated by specific number of predictors in USGS water management regions 1 to 9.

Approximately half of the regions are only dominated by three or less predictors model. Four and five predictor models are govering less that 10 percent of the stations in all different regions.

The spatial distribution of number of predictors are similar between flash flood duration and magnitude with very minimal differences. For example, the dominancy of the three or less predictors are more visible for intesity comparing to duration. In region 9, comparing the results from magnitude with duration, the four predictors model own more percentage of the stations for the duration than the one for the magnitude.

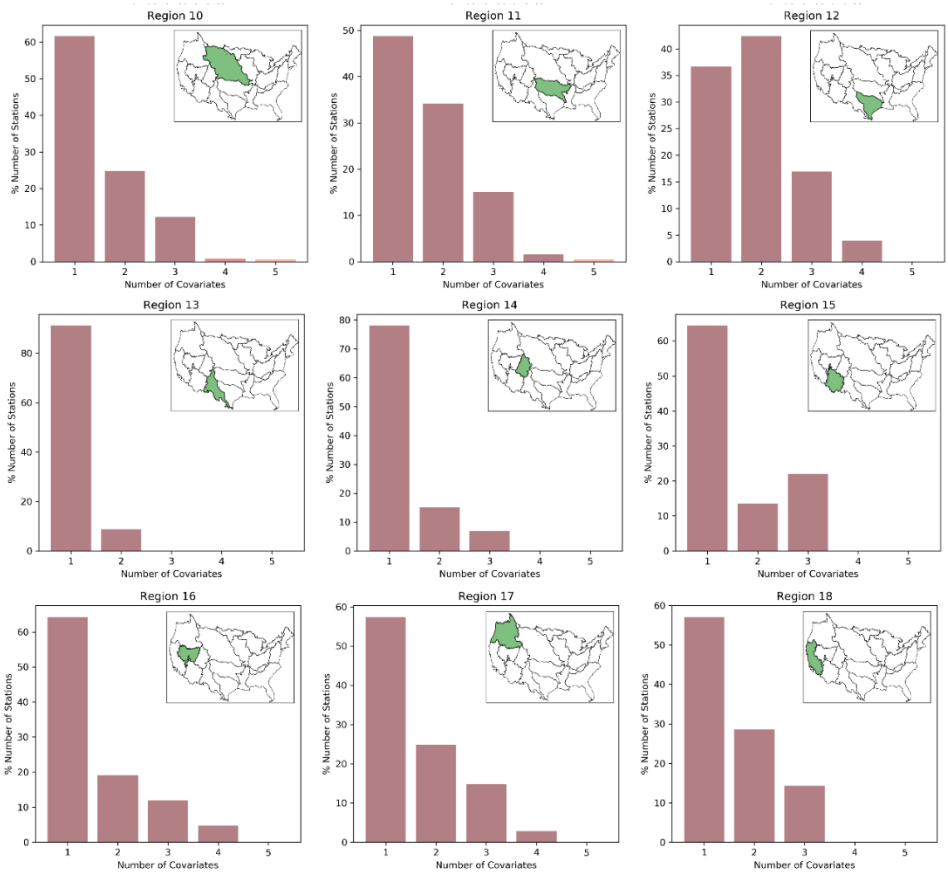

Figure 3-17. Percent of stations dominated by specific number of predictors in USGS water management regions 10 to 18.

Figures 3-18 and 3-19 show the percentage of stations in which each specific hydroclimatic variable has been chosen as one of the influential variables affecting the flash flood magnitude for each water resources region. Precipitation amount owns the highest percentage of stations in majority of the regions. In regions 9, 13, 14, and 15, precipitation amount is among the influential variables for flash flood magnitude in more that 70 percent of stations. Whereas, in the eastern regions, it covers around 30 to 40 percent of the stations. Precipitation rate does not show strong effect on flash flood magnitude with a highest rate of 30% in the eastern regions and approximately 10% in the western regions. Soil moisture is more effective in the eastern regions comparing to the

western ones. In the flash flood duration section, it was shown that soil moisture dominates region 8 and it overall has higher percentages comparing to the results for flash flood magnitude. Storm runoff indicates similarities between the results of flash flood duration and magnitude.

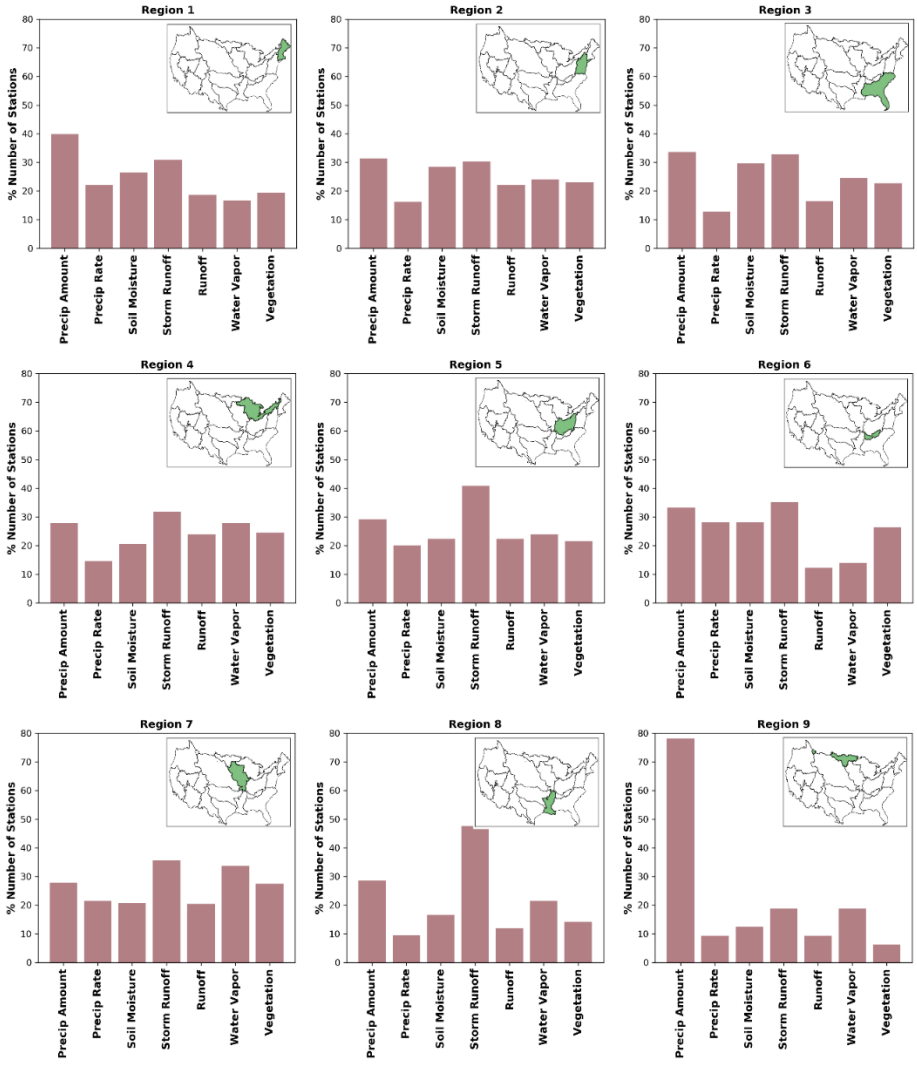

Figure 3-18. Percentage of stations dominated by specific number of predictors in USGS water management regions 1 to 9.

Storm runoff is the second most selected variable after precipitation amount in majority of the regions. It dominates the flash flood magnitude in regions 5, 6, and 12.

Runoff, as an influencing variable, varies between 10 to 30 percent coverage in the eastern regions and less that 20 percent for the western ones. Precipitable water has a constant governing rate of 20%, except in region 13, where it is not among the influencing variables. Vegetation has more influence on the duration of flash flood comparing to the magnitude. It has an approximately 20 to 30 percent influence on the duration, whereas this rate reduces to less than 10 percent for flash flood magnitude, specifically in the western regions. Overall, an extreme pattern is visible in the western regions where precipitation dominates the flash flood characteristics, whereas in the eastern regions, the variables show approximately equal contribution.

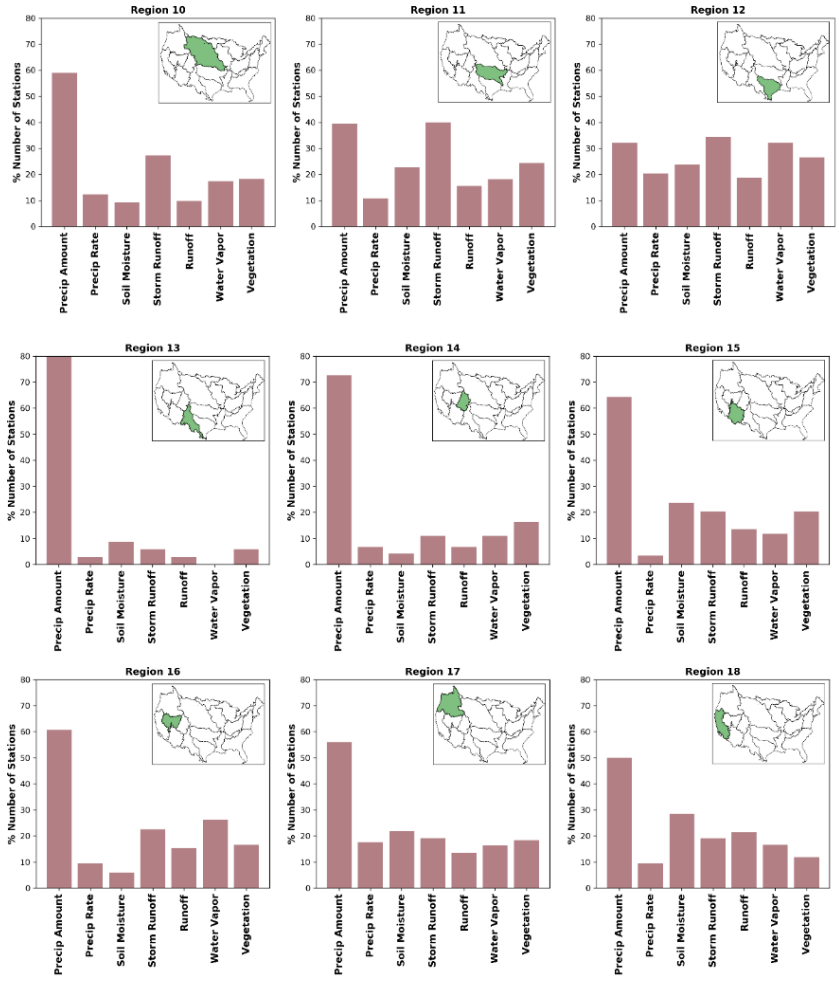

Figure 3-19. Percentage of stations dominated by specific number of predictors in USGS water management regions 10 to 18.

3.7 Summary and Conclusion

This chapter attempted to introduce a comprehensive methodology for assessing the effect of hydroclimatic variables (i.e., precipitation amount, precipitation rate, soil moisture, precipitable water, storm water runoff, runoff, and vegetation) on flash flood characteristics such as duration and magnitude. D-vine copula regression model was used to achieve this goal. Variables were added to the model sequentially to choose the most influencing variables with a new step-wise algorithm. The methodology was implemented

for 2,751 USGS stations over the CONUS. The spatial distribution of the chosen variables and number of predictors in the model were assessed both at the station level and 18 water resources regions defined by the USGS. The main findings of the study are as follows:

- Studying the correlation between the hydroclimatic variables and flash flood characteristics proves the need for a more complex approach to better understand this relation.
- The one-predictor models are governing the majority of stations over the United States for flash flood duration and magnitude, which is followed by the two-predictor models. Although seven predictors were fed to the model, the algorithm does not capture any significant improvement for implementing more than 5 predictors.
- Precipitation amount dominates among all the variables for its influence on flash flood magnitude and duration; whereas, the precipitable water and vegetation own the least contribution.
- Vegetation and soil moisture affect flash flood duration in more stations comparing to the flash flood magnitude.
- Western regions show a high percentage of the dependency on the precipitation amount comparing to the eastern regions among the hydroclimatic variables.

4 Modeling the Joint Influence of Hydroclimatic Variables on Flash Flood Characteristics using D-Vine Quantile Regression Model

4.1 Background

Identifying the occurrence of flash flooding, estimating the associated risk, and implanting effective mitigation measures require high spatial resolution flash flood forecasting with adequate lead-time. Several types of models have been used to forecast flash flooding, including empirical models and biophysical simulation models. Empirical and biophysical models are based on mathematical functions that simulate hydrologic processes at the basin scale (Singh and Woolhiser 2002). The shortcomings from the physically based models comes from the complexity, and time consuming and difficult calibration process.

As an alternative to physical-based models, data- driven modeling approach allows rapid construction of complex models to estimate outcomes based on historical experiences and events (Zhao et al. 2018). Data-driven models employ the statistical relation between the concurrent input and output time series to build a predictive framework and can be applied either alone (using a purely statistical model) or in conjunction with physics-based models, which use mathematical equations derived from the physical processes (creating a hybrid model) (Abrahart et al. 2007; Solomatine and Ostfeld 2008; Wu et al. 2009).

This chapter attempts to build a statistical flash flood forecasting system upon the methodology used in the previous chapter (i.e., D-Vine Copula quantile regression model). The main advantage of such technique is its capability in estimating uncertainty, which is one of the focuses in this study. The primary contribution of this research is to establish

and validate the suitability of a copula-statistical methodology for forecasting flash flood based on climate variables' influences.

4.2 Methodology

This study aims to forecast the conditional quantile q_a of the response variable Y (i.e, flash flood events) at some arbitrary level $a \in (0,1)$ for the given covariates $(X_1, \dots, X_n)$ (i.e., the chosen hydroclimatic variables). For this purpose, Kraus and Czado (2017) employed the inverse of conditional distribution based on D-vine copula as follows:

$$q_a(X_1, \dots, X_n) = F^{-1}_{Y|X_1,..,X_n}\langle a|x_1, \dots, x_n\rangle \quad \textbf{(4-1)}$$

By using the probability integral transform, whereas $V = F_Y(Y)$ and $V = F_n(X_n)$, the right hand side of Equation (4-1) can be formulated as:

$$F_{Y|X_1,..,X_n}\langle y|x_1, \dots, x_n\rangle = C_{V|U_1,..,U_n}\langle v|u_1, \dots, u_n\rangle \quad \textbf{(4-2)}$$

$$F^{-1}_{Y|X_1,..,X_n}\langle y|x_1, \dots, x_n\rangle = C^{-1}_{V|U_1,..,U_n}\langle v|u_1, \dots, u_n\rangle \quad \textbf{(4-3)}$$

Therefore, the estimate of conditional quantile function can be obtained by estimating the marginal F_Y and F_j, j=1,…,n. In section 3.5, the vine theorem and the approach to model the conditional quantiles were described. The model predictors and study time are similar to chapter 3. The predicted values for flash flood duration and magnitude will be evaluated at 10^{th}, 50^{th}, and 90^{th} quantiles.

4.3 Results and Discussion

This study implements the D-Vine copula quantile regression to find the uncertainty bound for flash flood characteristics (i.e., duration and intensity) based on the influential variables chosen in chapter 3. In the first section, the predictive power of the model is

evaluated against the observation for the median of flash flood magnitude and duration for the recorded events at USGS stations. Then, the success rate for each model is assessed at both the station level and regional scale. Finally, the success rates combined with the results from the previous chapter are examined to study the predictive power of the models at different regions.

4.3.1 Predictive Power of D-Vine Copula Quantile Regression Model

Figures 4-1 and 4-2 represent the difference between the model predictors and observation for flash flood duration and magnitude, respectively. For a better comparison, the median of flash flood characteristic at any given station is calculated for given USGS stations. The stations with less than 10 events are omitted for this part to ensure a meaningful statistical comparison.

Figure 4-1 shows the observed and modeled flash flood duration, and the absolute bias of the modeled results. The flash flood duration is the difference between the time at which flow exceeded the warning discharge threshold and the peak time (in hours). Results indicate that the highest flashiness events (i.e., longer duration) are mostly observed in Florida . The eastern part faces flash flood with less than 20 hours duration. However, there are several stations with high duration in the Midwest and southeast regions. Majority of the stations in the Western United States face lower flash flood durations.

The predicted flash flood duration based on the D-Vine Copula quantile regression is shown in the middle map in figure 4-1. The modeled duration is following similar spatial pattern comparing to the observed one. To have a clearer comparison, the absolute difference between the median of observed and predicted flash flood duration is calculated

and shown in the bottom graph. More than 90% of stations indicate an absolute bias less than 2 hours. Among 2,751 considered stations, less than 20 station demonstrate over 15 hours difference between modeled and observed flash flood duration. This can be due to the model fit and more complexity in those regions, which can be further investigated in future studies.

Figure 4-2 shows the median result for the observed and predicted flash flood magnitude. The modeled intensities show promising outcomes. The majority of stations in the West coast region demonstrate flash flood with approximately 1000 m^2/s magnitude. Whereas, the ones in the central and eastern U.S. have lower intensities. A cluster of stations with high magnitude of flash flood are located in the lower parts of Midwest and North East.

The absolute bias between the observed and predicted flash flood magnitude ranges between 0 to 50 m^2/s, with majority of them ranging less than 10 m^2/s. Unlike the predicted duration, the higher magnitude of absolute bias are noticeable in the Western and Eastern regions. Whereas the central U.S. owns the highest accuracy for the predicted values. In more detail, the stations with the highest flash flood magnitudes are located in these regions. As it is addressed in different studies, predicting the extreme values are more challenging comparing to the moderate or low flow events (Collier 2007; Gourley et al. 2010; Douinot et al. 2016; Hardy et al. 2016).

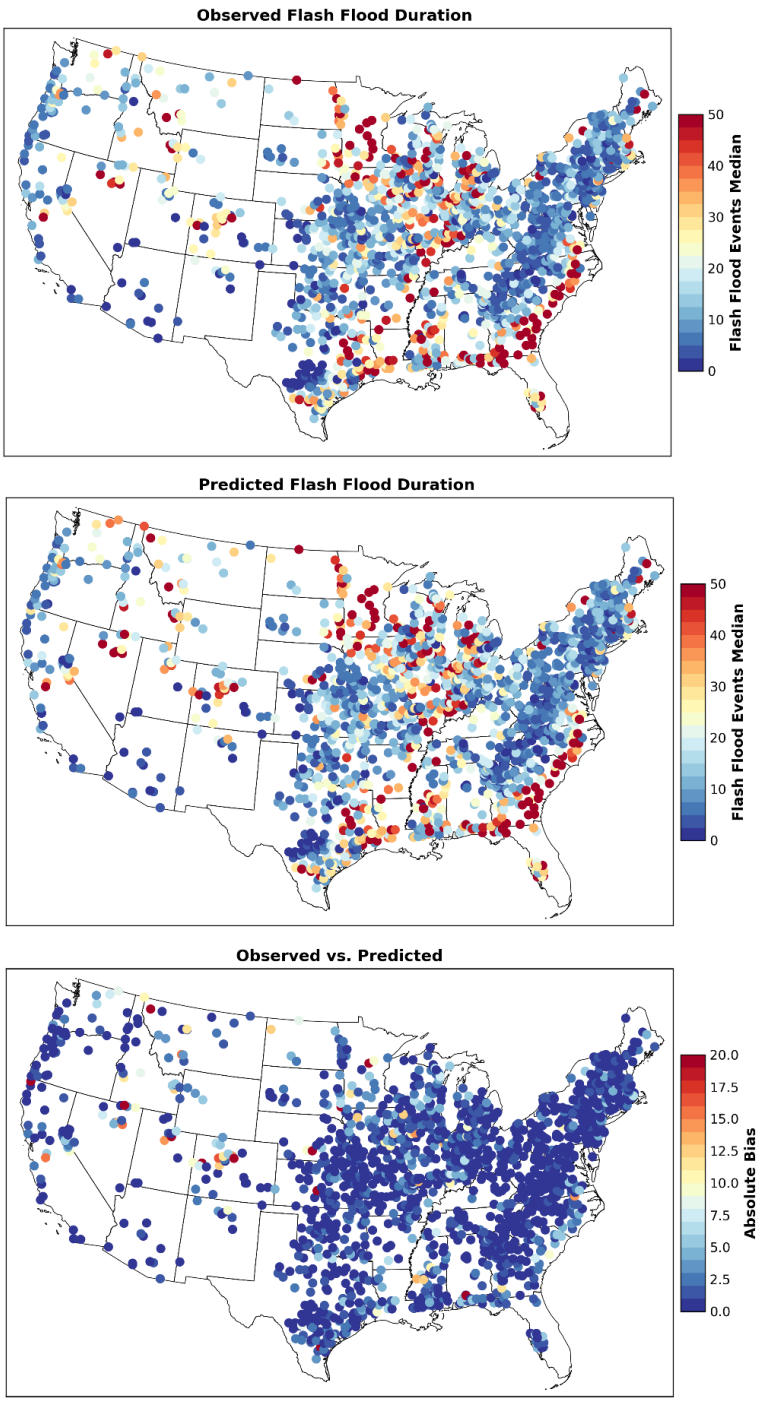

Figure 4-1. Median of Flash Flood Duration calculated for Observation (Top) and simulation (Middle), and the Absolute Difference between Observation and Modeled Flash Flood Duration (Bottom), all shown in hours.

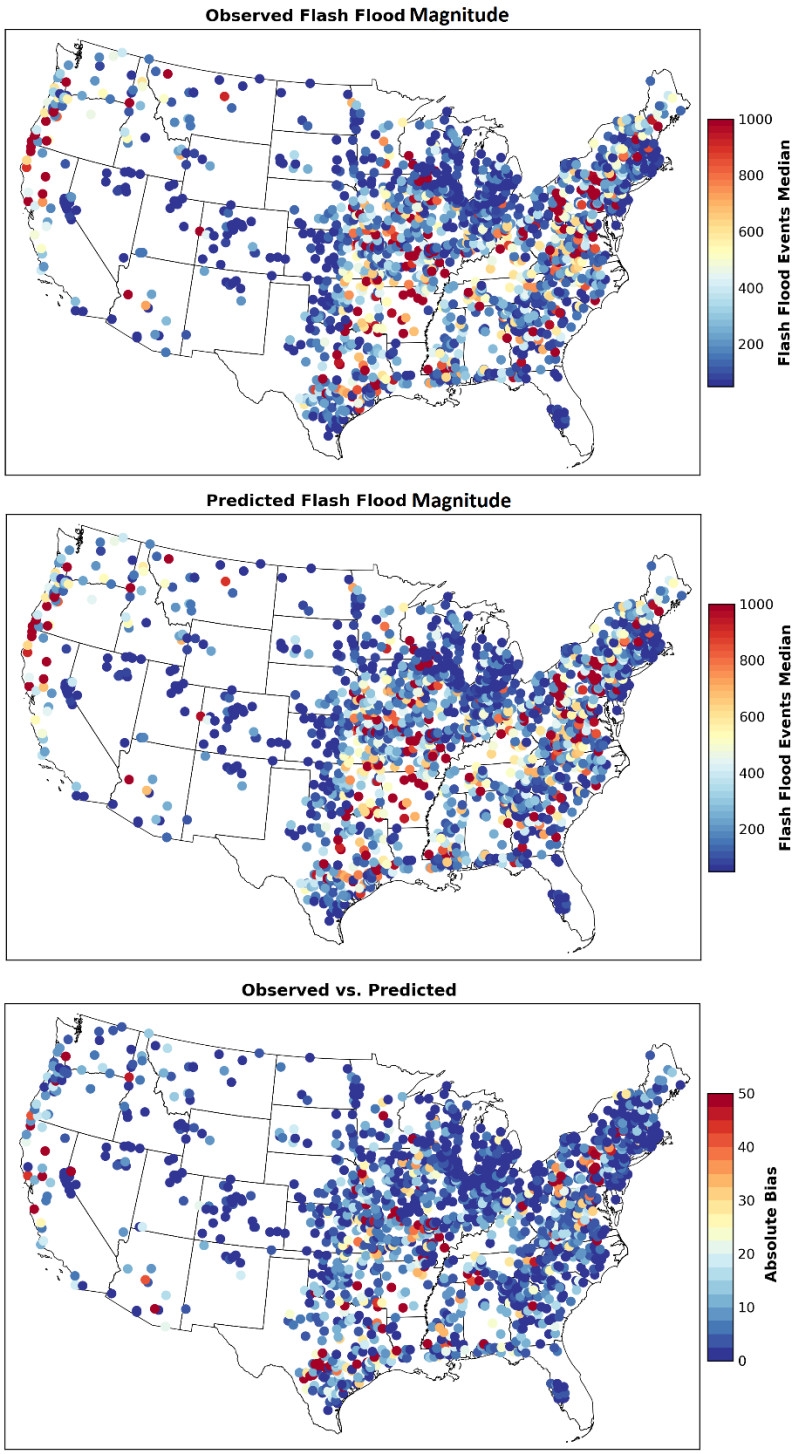

Figure 4-2. Flash Flood magnitude Median calculated for Observation (Top) and Modeled (Middle), and the Absolute Difference between Observation and Modeled Flash Flood Duration (Bottom).

4.3.2 Flash Flood Prediction System Success Rate

To evaluate the predictive power of the modeled quantiles (10th and 90th) from the D-vine copula model, the success rate is calculated at each USGS station. Success rate of a given station is defined as the rate that the observed flash flood duration/magnitude falls within the predicted range of 10-90th modeled quantiles during the study period.

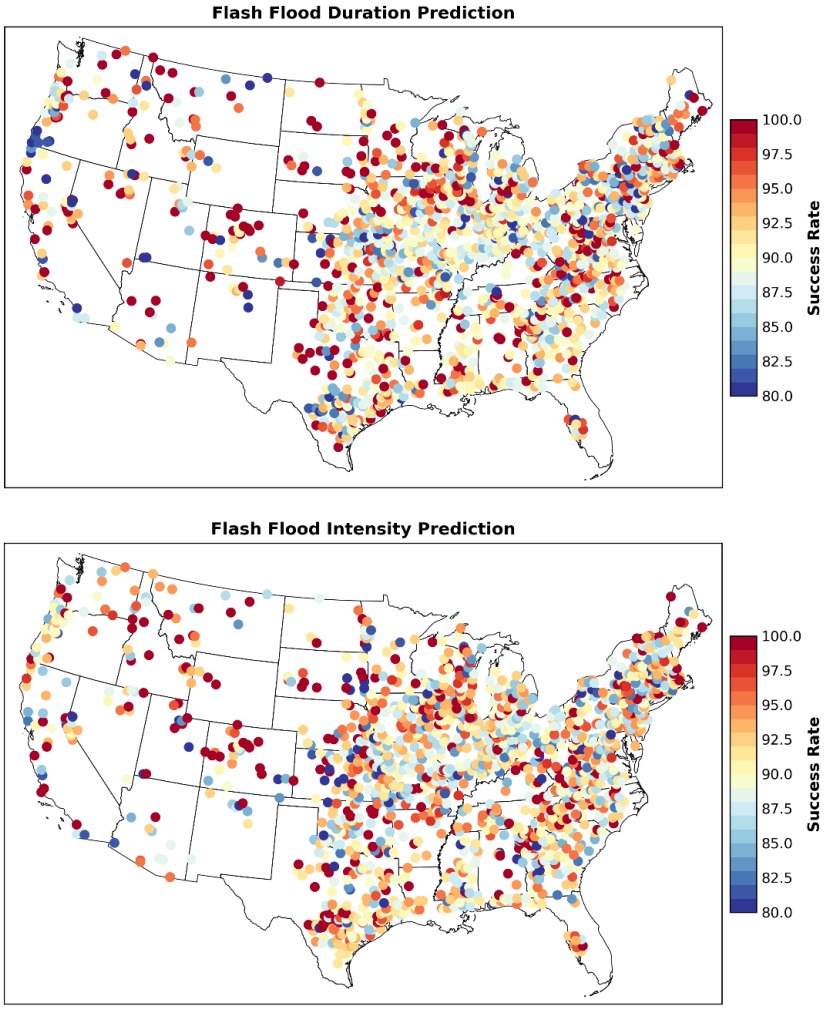

Figure 4-3. Success rate for the modeled flash flood duration (top) and magnitude (bottom) using the D-Vine Copula Regression model.

Figure 4-3 demonstrates the success rate over the United States for predicted flash flood duration (top) and magnitude (bottom). Mojority of the stations are showing success rate of more than 85 percent. There are a significant number of stations with 100% success rate, which indicates that the model is successful to predict the probable bound for flash flood characteristics in those stations. In less than 10% of stations, the success rate is less than 80%, which may be improved by taking more variables into account or longer period of time in training the model.

Figure 4-5 and 4-6 present the regional distribution of success rate for flash flood magnitude and duration (blue and mustard violin plots, respectively) along with the number of stations at each range (represented using white dots). The color used for each region is presented in Figure 4-4.

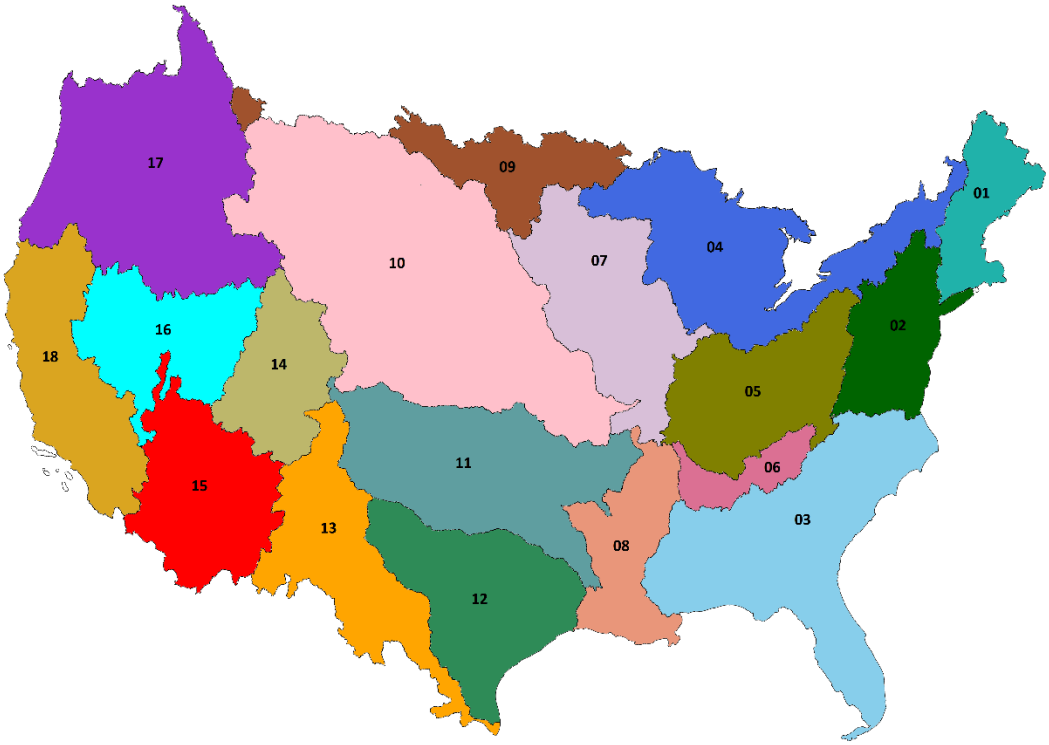

Figure 4-4. USGS Water Resources Regions. The colors used for each region is similarly utilized for plotting Figures 4-5 and 4-6.

Most of Western regions including regions 13 to 18 as well as region 9 own fewer stations, and among those station, the majority have less than 10 recorded flash flood events in 30 years of study period. Whereas, the concentration of highly impacted stations increases moving eastward where regions 1 to 7 are located. Distribution of success rate is close to uniform in region 1, 6 and 8 for predicted duration. Whereas the distribution of the rest of regions are skewed to the right and higher success rates, except for region 9. Flat shaped distribution is observed for predicted magnitude in regions 6, 9, and 18. However, the rest of the regions show skewness toward higher success rate distribution for the modeled flash flood magnitude.

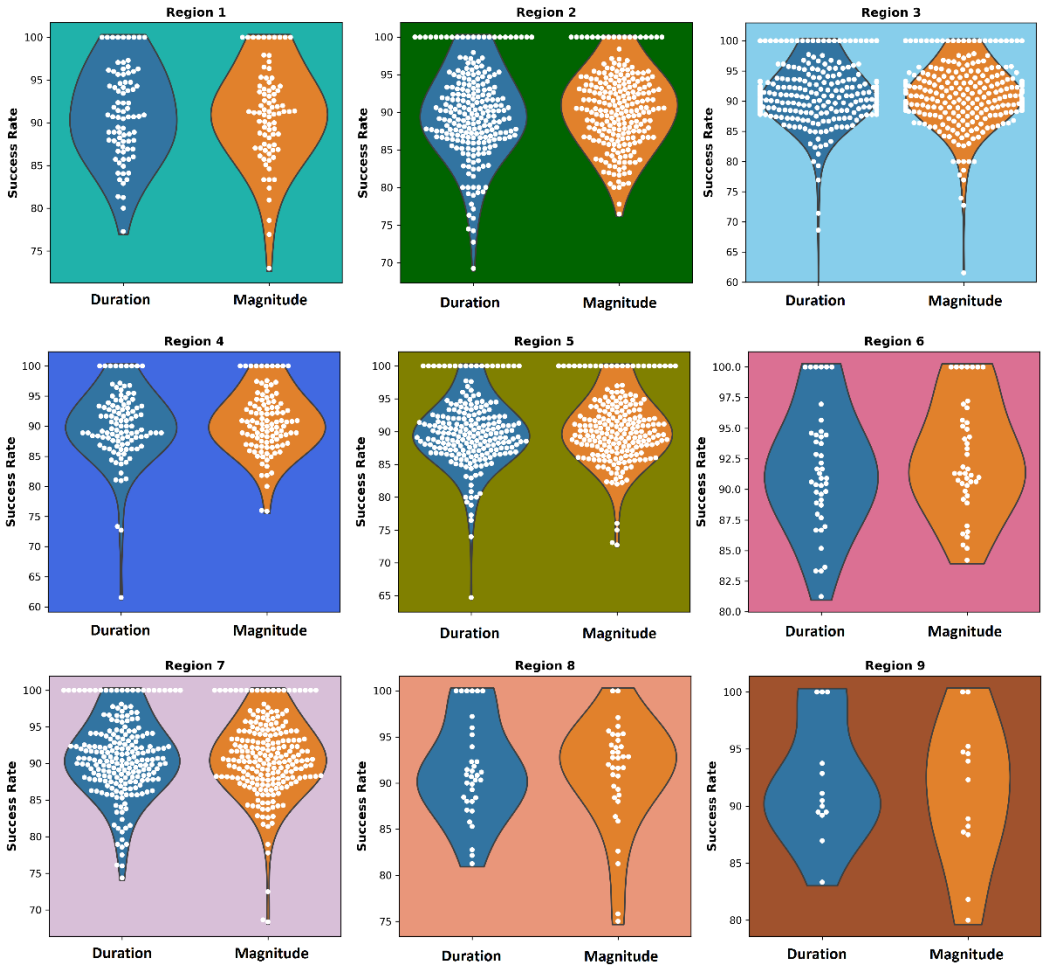

Figure 4-5. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model for the flash flood duration and magnitude for regions 1 to 9.

Stations located in the central and eastern parts (regions 2, 3, 5, and 7) and the central regions (10, 11, and 12) own more stations with a 100% success rate. These regions are occupied by stations with high flash flood magnitude and duration (Figures 4-1 and 4-2). Region 3 possesses stations with highest flash flood durations, the median success rate of which is approximately 90%, and it owns high concentration of stations with more than 90% success rate. Overall, all the regions have a median success rate of 85% to 90%, with less than 3 stations indicating a success rate of less than 75% (regions 11, 12, and 17). The

range of success rate is higher for predicting flash flood magnitude comparing to the duration. Furthermore, the majority of variables used in the model have direct impact on flash flood magnitude in comparison to the duration.

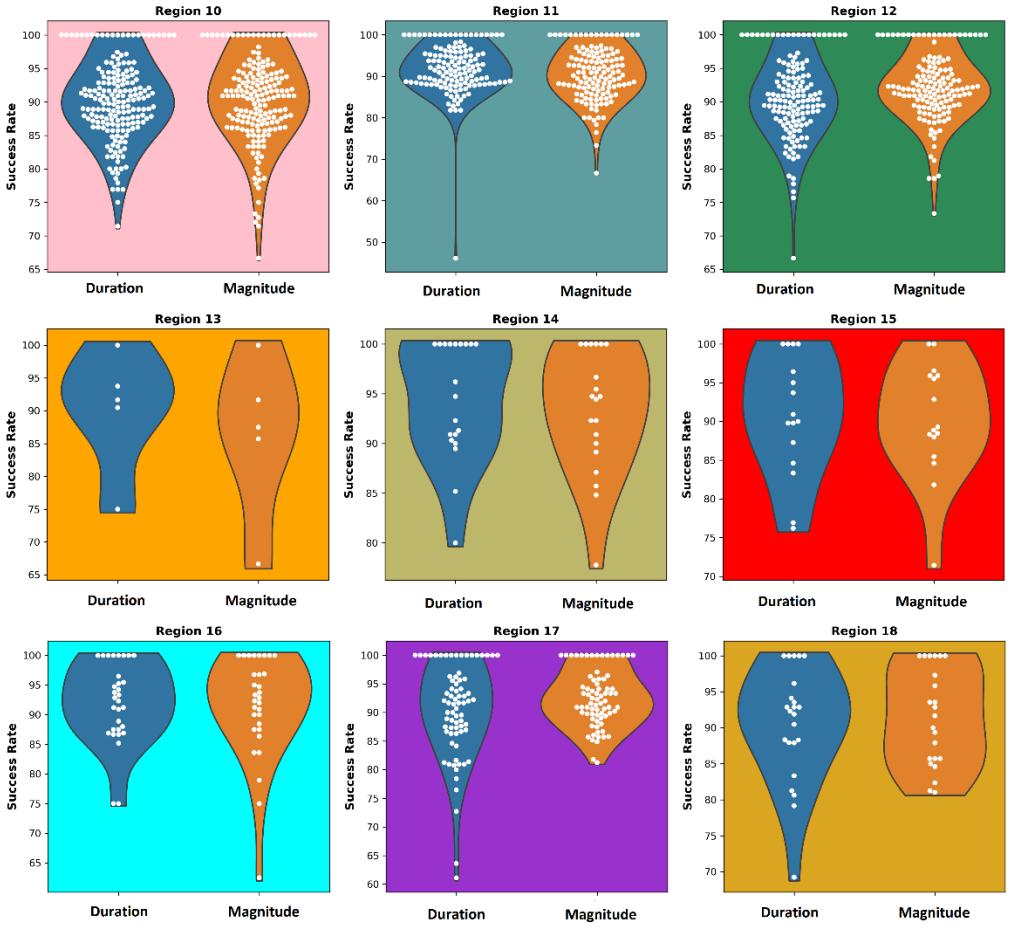

Figure 4-6. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model for the flash flood duration and magnitude for region 10 to 18.

4.3.3 Success Rate vs. Number of Predictors

To evaluate the predictive power of the D-Vine copula model with different number of predictors, the stations are categorized into five classes based on the number of predictors. Figures 4-7 and 4-8 show the success rate for the stations in these five categories at the 18 water resources regions for flash flood duration. One predictor models show a range of 80% to 100% for success rate in the majority of regions. In more that 50 percent of the regions, the one predictor models show a considerable number of stations with 100% success rate. Whereas the regions with few stations show less success rate for one predictor models including the Southwestern regions and Colorado (regions 13 to 15).

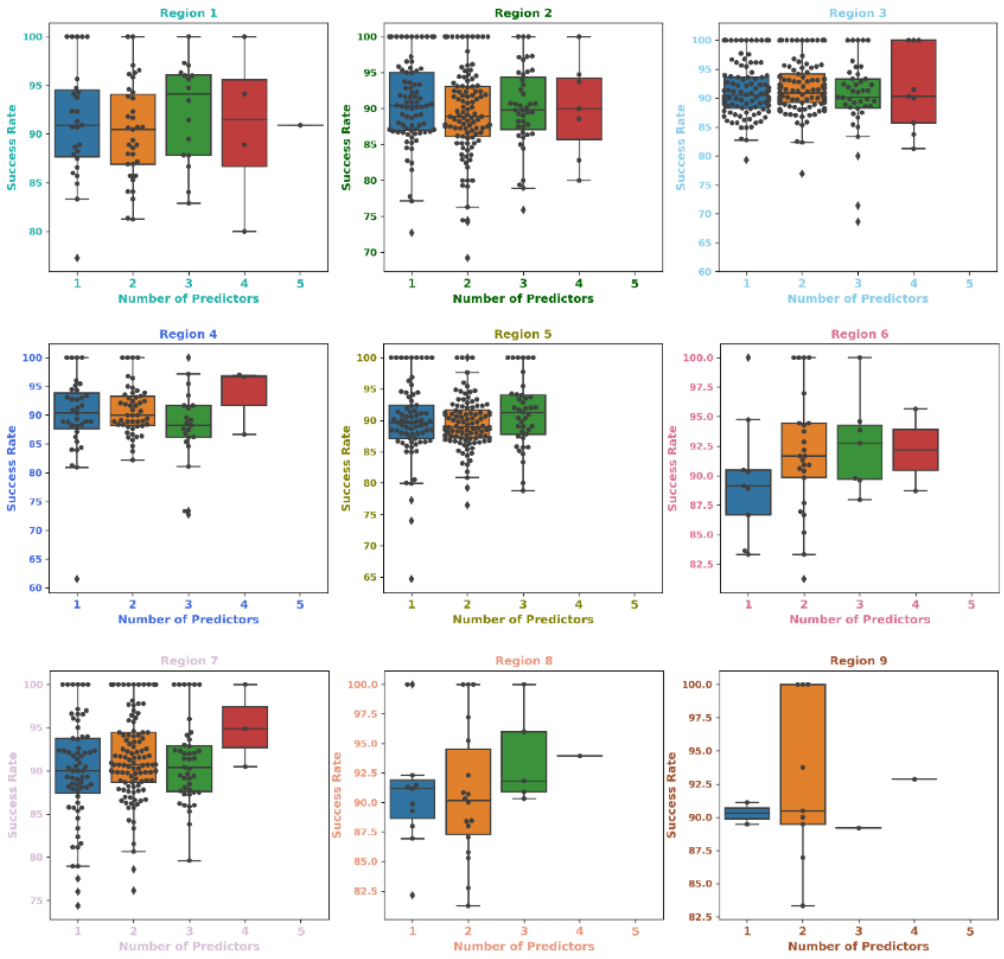

Figure 4-7. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the number of predictors for the flash flood duration including regions 1 to 9.

Regions 9 and 13 retain the lowest number of two-predictor stations with more than 10 flash flood events during the study period. However, the predictive power of the model is relatively high for these stations. The two-predictor models show similar or higher accuracy comparing to the one predictor models, even though the number of stations for the former are less than that for the latter (i.e. one predictor models). The median success rate for the two-predictor models is higher than 90%. Whereas, it drops to 80% in some regions for one-predictor models. The same pattern is observed for the three-predictor

models in the eastern regions comparing to the one- and two–predictor models. However, the success rate drops to less than 85% in California (region 18). Seven regions do not possess any stations with four predictors. In the regions that four predictors are present, they have equal or higher success rate comparing to lower dimensions expect in Pacific Northwest (region 17).

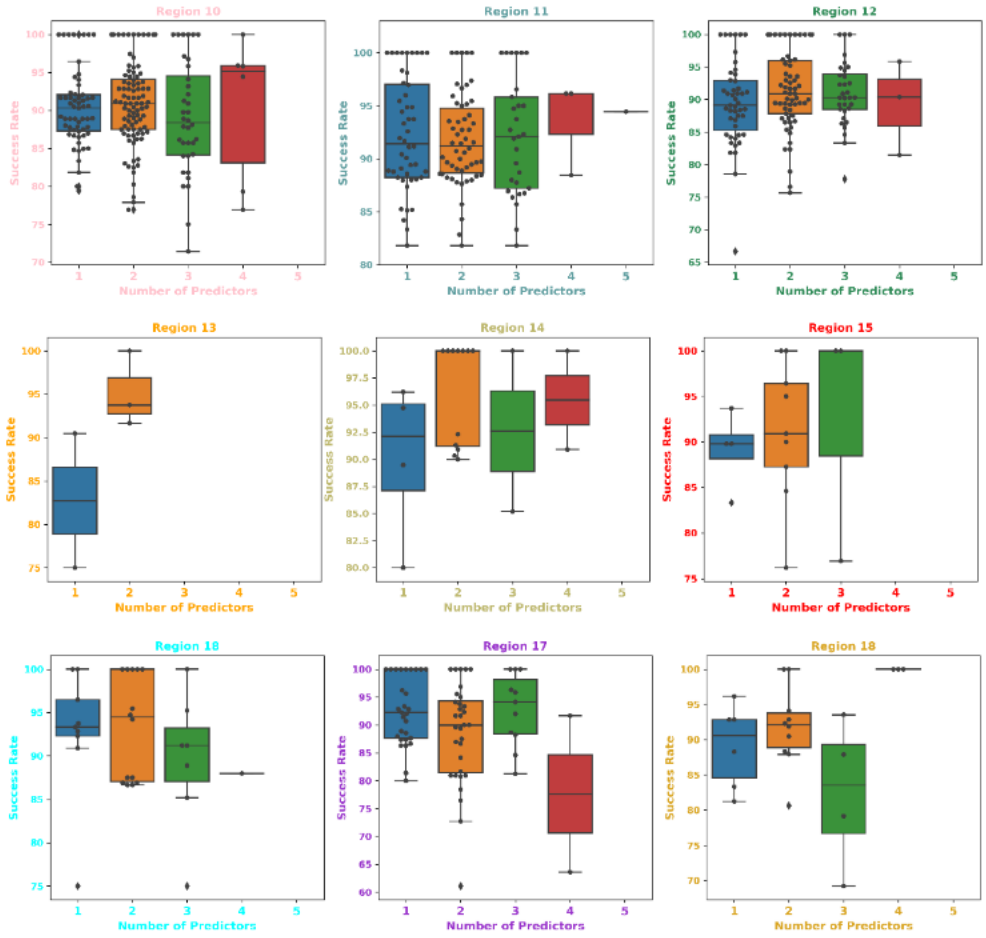

Figure 4-8. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the number of predictors for the flash flood duration including regions 10 to 18.

Figures 4-9 and 4-10 show the success rate for the stations in five categories of number of predictors at the 18 water resources regions for flash flood magnitude. The higher accuracy of the two-predictor models is noticeable in most of the regions except for region 13 which has 5 stations. The three-predictor models demonstrate higher success rate in the eastern regions in comparison to the western United States. The four-predictor models are selected in ten regions and due to their low number of stations, it is not accurate to conclude on the prediction power comparing to the rest of predictor combinations.

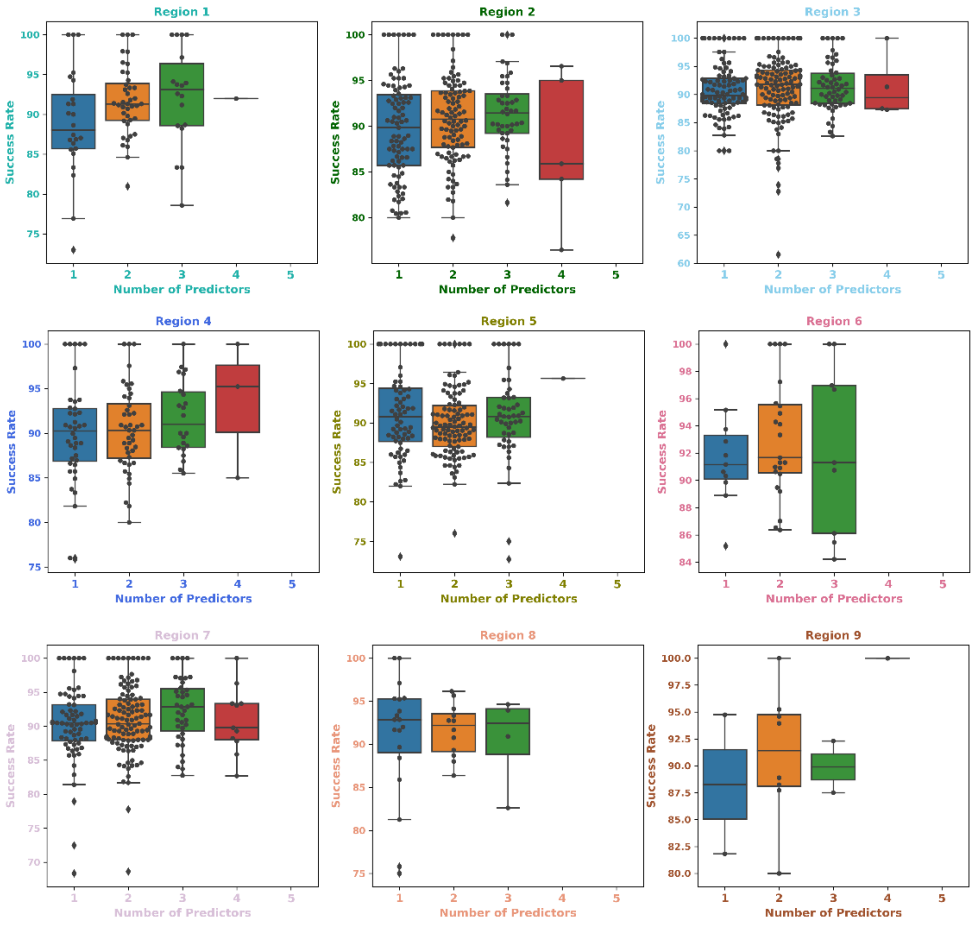

Figure 4-9. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the number of predictors for the flash flood magnitude including regions 1 to 9.

The five-predictor models are visible in region 10 with the lowest success rate, this can be explained by the complex structure of the model in higher dimensions (Kraus and Czado 2017). Overall, comparing the relation between success rate and number of predictors, the two- and three-predictor models generally have higher success rates comparing to the one-predictor models. This pattern shows that the combinations of hydroclimatic variables can enhance the accuracy of the flash flood prediction system.

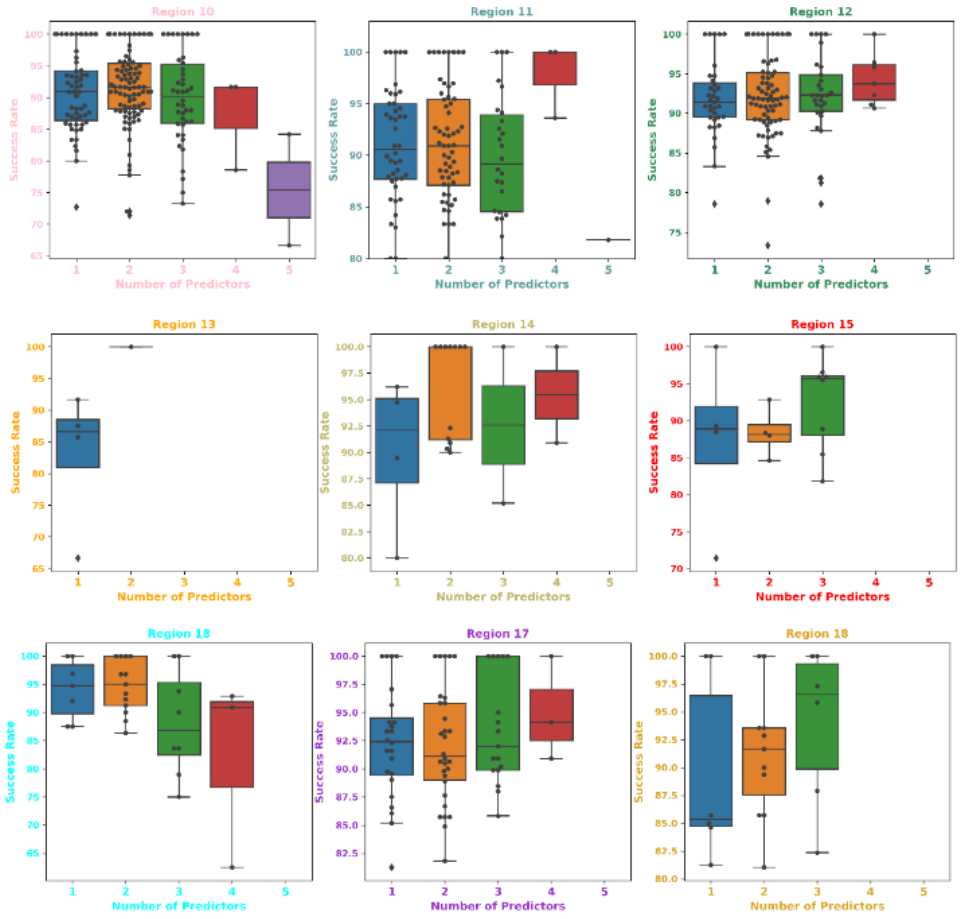

Figure 4-10. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the number of predictors for the flash flood magnitude including region 10 to 18.

4.3.4 Success Rate vs. Predictors

Figures 4-11 and 4-12 present the success rate of the proposed model for flash flood duration in comparison to the hydroclimatic variables involved in the model. The models that utilized precipitation amount as a predictor retain a fairly high success rate in majority of the regions. Region 13 , which is located in the southwest region and does not receive a lot of precipitation, does not include precipitation amount as a predictor of flash flood.

Precipitation rate owns equal or higher success rates in the eastern regions comparing to the precipitation amount. Soil moisture indicates much steadier pattern comparing to the rest of variables, its influence is observable in majority of the regions except region 13, and it has a fairly high range of success rate between 80% to 100%. Storm runoff success rate is higher than the runoff success rate in the eastern regions; whereas, it is the opposite case for the western regions. This proves the dominancy of precipitation amount in the eastern regions in controlling the flash flood duration. Water vapor has a similar pattern to precipitation amount, which shows its effect on forming precipitation. In addition, in some regions such as regions 6, 7, and 8, water vapor shows higher success rates in comparison to the rest of the predictors.

The difference between the success rates of each predictor is more noticeable in the western regions in comparison to the eastern ones. Central regions, specifically region 11, yields very high success rates for all the hydroclimatic variables included in the model. The variance of success rate among the variables increases as the number of stations decreases in regions. The dramatic patterns are observed in regions 9, 13, and 14, where the lowest

clustering of stations exists and they receive infrequent flash flooding compared to other regions.

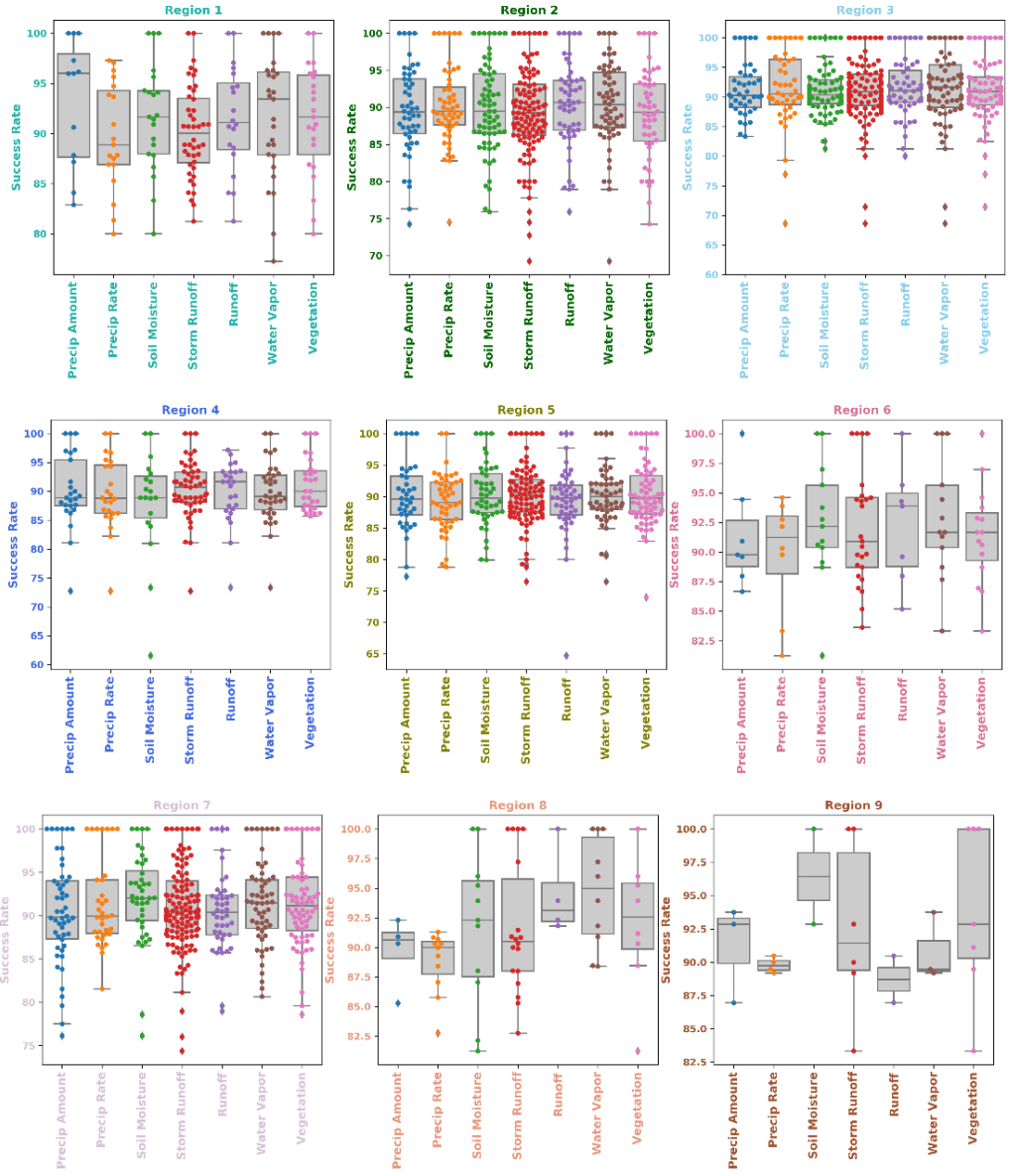

Figure 4-11. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the predictors involved for modeling flash flood duration in regions 1 to 9.

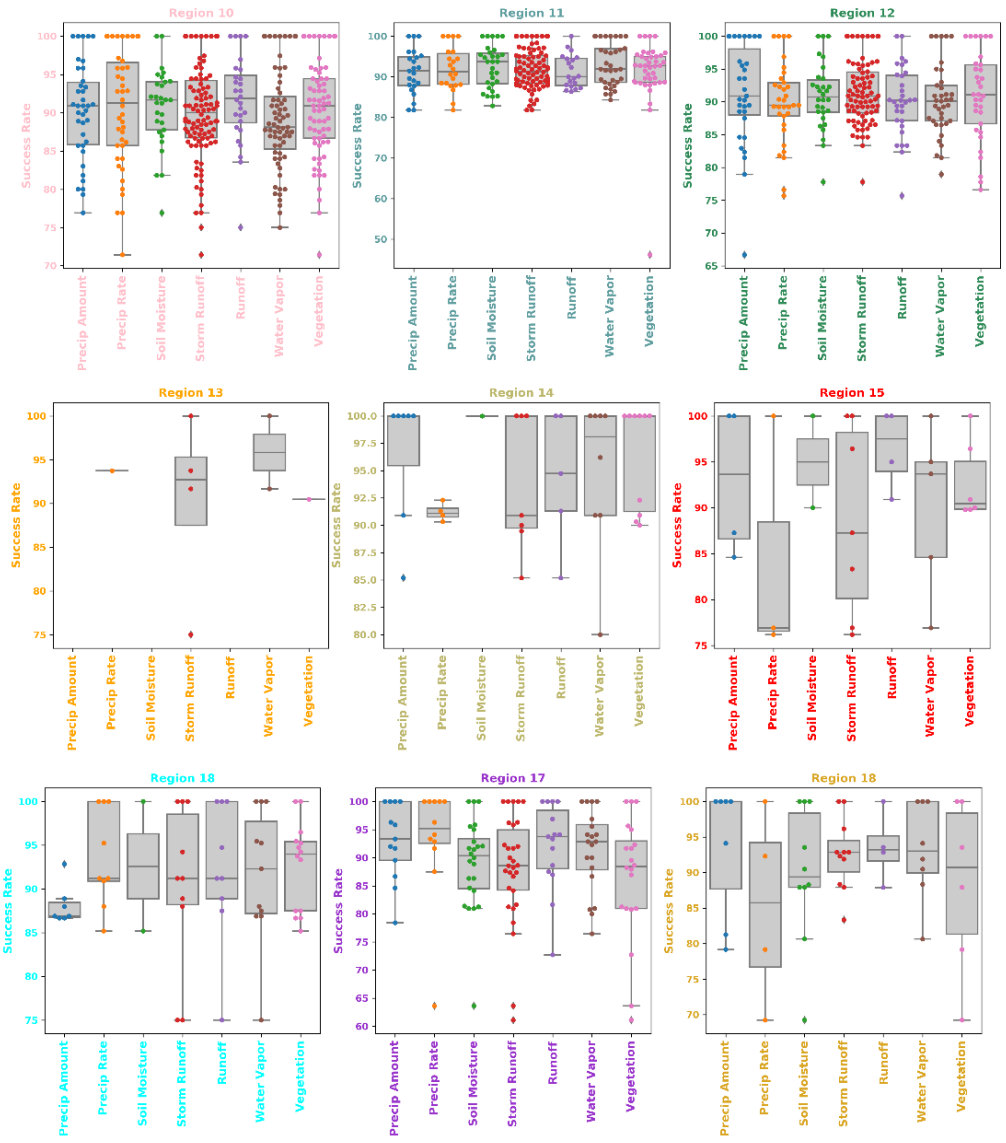

Figure 4-12. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the predictors for the flash flood duration including region 10 to 18.

Figures 4-13 and 4-14 present the results for the success rate of predictors for modeling flash flood magnitude. The regional pattern for the results of flash flood magnitude is closely similar to the results of duration. The eastern regions retain similar success ranges for all the predictors in the model. Whereas, the fluctuation is sensible in

the western regions. However, the success rates for all the hydroclimatic variables in modeling flash flood magnitude are not as high as that for flash flood duration in region 11. The success rates for precipitation amount is higher than that for precipitation rate in the eastern regions; whereas, it shows lower rates in the Western regions. This proves that in the eastern regions, flash flood magnitude is more affected by precipitation amount; however, the models involving precipitation rate show more precision in the western regions.

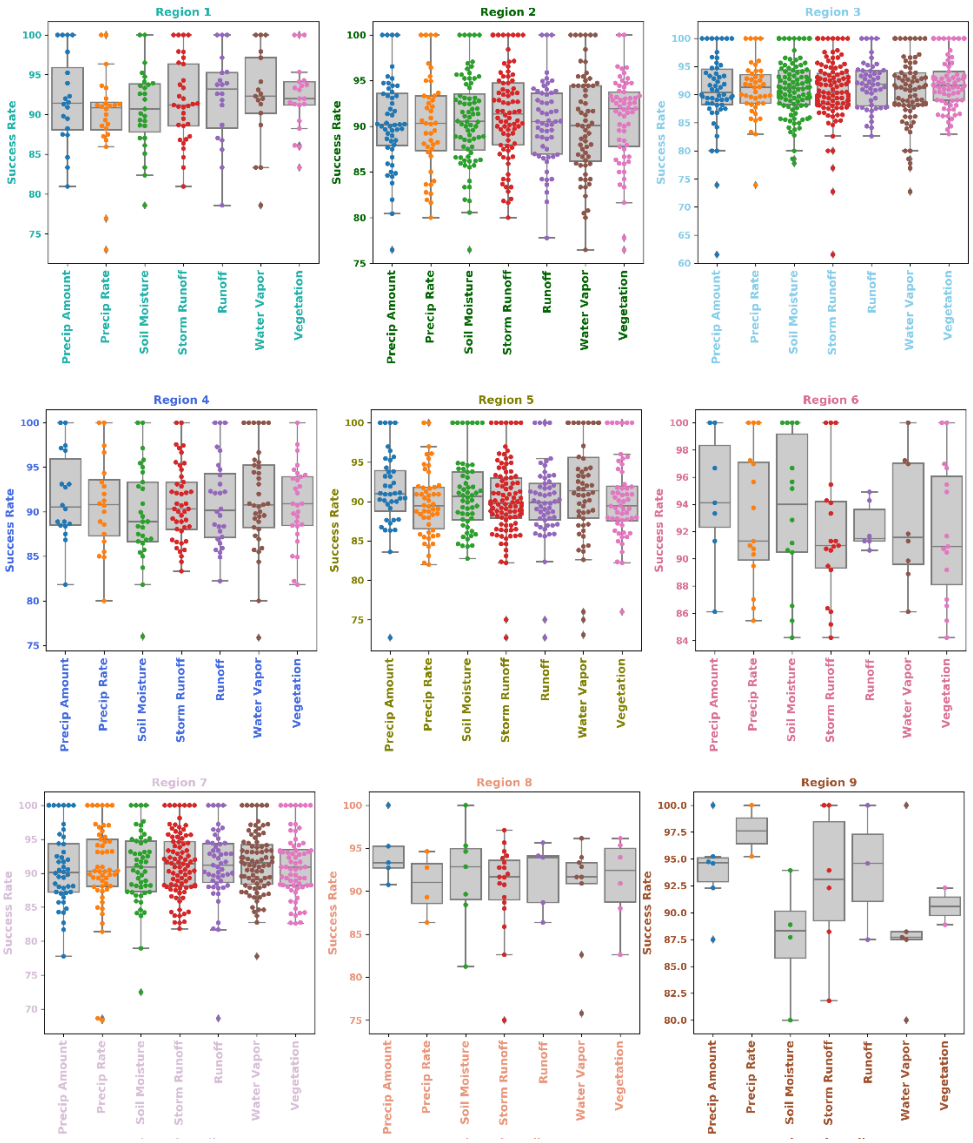

Figure 4-13. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the predictors for the flash flood magnitude including region 1 to 9.

The stations where they consider storm water runoff in their predictive models are more successful in predicting the flash flood magnitude comparing to the ones that involve runoff in the Eastern regions. This behavior is the opposite in the Western regions. There is a dramatic difference between predictors' success rate in California and Pacific region

(regions 17 and 18). These two regions own the highest success rates for all the variables among the Western regions.

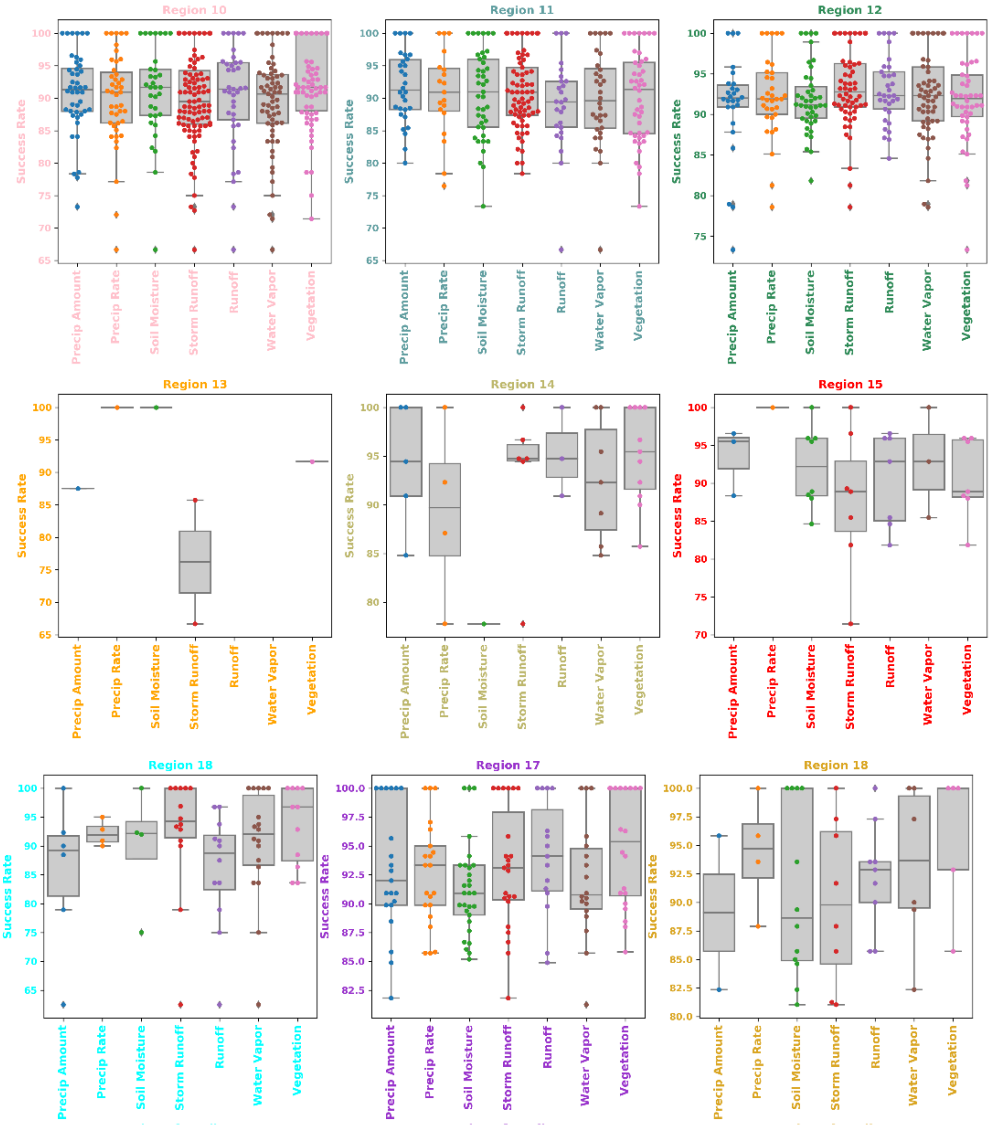

Figure 4-14. Success rate for the predicted uncertainty bound calculated by D-Vine Copula Regression model vs. the predictors for the flash flood magnitude including region 10 to 18.

4.4 Summary and conclusion

This study extended and employed the model that was developed in Chapter 3 to predict flash flood characteristics (i.e., duration and magnitude) considering the underlying uncertainties. The D-Vine quantile regression model was used to estimate the 10^{th} and 90^{th} quantile of flash flood magnitude and duration. The model results were compared with observation in each station for flash flood duration and magnitude. Furthermore, to evaluate the predicted uncertainty, the success rate was calculated as the rate that an observed event falls between the predictive quantiles. To extend the evaluation, the model predictive power was assessed based on the number of predictors as well as the hydroclimatic variables that were included as predictors in the model. The main findings of this chapter are as follows:

- The spatial pattern of the stations' median for the flash flood duration and magnitude is successfully replicated by the modeled values comparing to the observation. The absolute bias is found to be within acceptable and promising range for both flash flood magnitude and duration.
- There are noticeable number of stations with success rate of 100%, specifically in the eastern regions of the U.S. The predicted flash flood magnitude owns higher success rates comparing to the modeled flash flood duration.
- The success rate is elevated as the number of predictors increases to three hydroclimatic variables. This pattern is more visible in the eastern regions comparing to the western ones.

- The success rate for different predictors in the models has less variation in the eastern regions comparing to the western region.
- The regions with fewer stations and infrequent flash flood events are more challenging for prediction.

5 Conclusion and Future Studies

In this dissertation, the main objectives included assessing the socio-economic vulnerability of the United States to flash flooding, and investigating the joint influence of hydroclimatic variables on Flash flood characteristics such as magnitude and duration. These objectives were accomplished by employing advanced statistical procedures.

This study can be divided into two phases. In the first phase, the focus was on comprehensive assessment of socio-economic vulnerability and its interaction with flash flood characteristics over the CONUS. Socio-economic vulnerability was assessed on both county and state levels by employing 36 social and economic variables. Probabilistic Principal Component Analysis (PPCA) was utilized to quantify socio-economic vulnerability index (SEVI). The flash flood characteristics including frequency, magnitude, duration, and severity were considered for this evaluation. Spatial distribution of flash floods was assessed using hotspot analysis to detect the clusters of high values and low values. The intersection of SEVI and flash flood characteristics were mapped using cross-tabulation. In addition, flash flood fatalities were employed for validating the calculated socio-economic index. Critical counties (i.e., high-vulnerable-hotspot) are clustered in the southern regions of the CONUS; whereas, the majority of non-critical counties (i.e., low-vulnerable-coldspot) are located in the Northern Great Plains.

The second phase of this study attempted to develop a statistical procedure for flash flood prediction based on the joint interaction of hydroclimatic variables. First, a comprehensive methodology for assessment of the influence of hydroclimatic variables (i.e., precipitation amount, precipitation rate, soil moisture, precipitable water, storm water

runoff, runoff, and vegetation) on flash flood characteristics, such as duration and magnitude, was introduced. D-vine copula quantile regression model was implemented to achieve this goal. The hydroclimatic variables were added to the model sequentially to choose the most influencing variables with a novel step-wise algorithm. The methodology was implemented for 2,751 USGS stations over the CONUS. The spatial distribution of the chosen variables and the number of predictors in the model were assessed both at the station level and across the 18 water resources regions defined by the USGS. Majority of the stations over the CONUS are governed by one predictor models, and precipitation amount is the dominating variable among the other examined hydroclimatic variables. In addition, the western regions show a high percentage of dependency on precipitation amount comparing to the eastern regions.

Finally, the D-vine copula quantile regression model was utilized to estimate the 10^{th} and 90^{th} quantile of flash flood magnitude and duration. The model results were compared with observation in each station for flash flood duration and magnitude. Furthermore, to evaluate the predictive power of the model, the success rate was calculated as the ratio that an observed event (i.e., flash flood duration and magnitude) falls between the predictive quantiles of the D-vine copula quantile regression model. To extend the evaluation, the model's predictive power was assessed based on the model complexity (i.e., number of predictors in the model) as well as the hydroclimatic variables that were considered as predictors. The model assessment demonstrated promising range for the absolute bias between the observation and predicted flash flood characteristics. Furthermore, there are noticeable number of stations with success rate of 100%,

specifically in the eastern regions of the U.S. Overall, the proposed technique indicates promising results that can be employed as an early warning system for flash flooding over the United States.

While comprehensive analyses were carried out to provide accurate and reliable assessments, this study can be further improved from various perspectives considering socio-economic vulnerability and forecasting system. Suggestions regarding improvements on each sector are explained in the following:

1. Socio-economic Vulnerability

The calculated socio-economic vulnerability index (SEVI) is based on only one point in time, a shortcoming that needs attention. The socio-economic status and flash flood potential inevitably change over time, due to climate change, migration, and social development. Although social vulnerability indices can efficiently describe broad-scale vulnerability, they may fall short in capturing more localized information related to exposure, sensitivity, and adaptive capacity that is often better collected using qualitative methods. Therefore, it would be beneficial to validate the calculated socio-economic vulnerability by comparing the post hazard outcomes with the pre-hazard vulnerability. A comprehensive assessment of the counties that have the potential to change their status from non-critical to critical is beneficial for future studies.

2. Flash Flood Forecasting System

This study chose a limited number of hydroclimatic variables including precipitation amount, precipitation rate, soil moisture, precipitable water, storm water

runoff, runoff, and vegetation to develop the data-driven flash flood forecasting system. A comprehensive study of the influential variables can improve the predictive power of the forecasting system, and variables such as urbanization and coastal characteristics for coastal areas may also be considered. Combining hydroclimatic variables from different sources can reduce the bias that an individual model (e.g., CFS) may introduce to the forecasting system. Flash floods are localized events, and thus, increasing the spatial resolution may result in more accurate and practical forecasting systems. Lastly, this newly developed technique can be adapted into an operational flash flood forecasting system combined with the socio-economic status of a certain location to predict the risk of flash flood events.

6 References

Abrahart RJ, See LM, Solomatine DP, Toth E (2007) Data-driven approaches, optimization and model integration: hydrological applications. Hydrol Earth Syst Sci 11:

Adger WN (2006) Vulnerability. Glob Environ Chang 16:268–281

Ahmadalipour A, Moradkhani H (2018) Multi-dimensional assessment of drought vulnerability in Africa: 1960–2100. Sci Total Environ 644:520–535

Aho K, Derryberry D, Peterson T (2014) Model selection for ecologists: the worldviews of AIC and BIC. Ecology 95:631–636

Armah FA, Yawson DO, Yengoh GT, et al (2010) Impact of floods on livelihoods and vulnerability of natural resource dependent communities in Northern Ghana. Water 2:120–139

Aroca-Jiménez E, Bodoque JM, García JA, Díez-Herrero A (2018) A quantitative methodology for the assessment of the regional economic vulnerability to flash floods. J Hydrol 565:386–399

Ashley ST, Ashley WS (2008) Flood fatalities in the United States. J Appl Meteorol Climatol 47:805–818

Bedford T, Cooke RM (2001) Probability density decomposition for conditionally dependent random variables modeled by vines. Ann Math Artif Intell 32:245–268

Bernard C, Czado C (2015) Conditional quantiles and tail dependence. J Multivar Anal 138:104–126

Binley A, Hubbard SS, Huisman JA, et al (2015) The emergence of hydrogeophysics for improved understanding of subsurface processes over multiple scales. Water Resour Res 51:3837–3866

Birkmann J, Cardona OD, Carreño ML, et al (2013) Framing vulnerability, risk and societal responses: the MOVE framework. Nat hazards 67:193–211

Blaikie P, Cannon T, Davis I, Wisner B (2014) At risk: natural hazards, people's vulnerability and disasters. Routledge

Borga M (2002) Accuracy of radar rainfall estimates for streamflow simulation. J Hydrol 267:26–39

Borga M, Anagnostou EN, Blöschl G, Creutin J-D (2011) Flash flood forecasting, warning and risk management: the HYDRATE project. Environ Sci Policy 14:834–844. doi: http://dx.doi.org/10.1016/j.envsci.2011.05.017

Braud I, Vincendon B, Anquetin S, et al (2018) 3 - The Challenges of Flash Flood Forecasting. In: Lutoff C, Durand SBT-MFHEE 1 (eds). Elsevier, pp 63–88

Brechmann E, Schepsmeier U (2013) Cdvine: Modeling dependence with c-and d-vine copulas in r. J Stat Softw 52:1–27

Center H (2002) Human links to coastal disasters. H John Heinz III Cent

Collier CG (2007) Flash flood forecasting: What are the limits of predictability? Q J R Meteorol Soc 133:3–23

Council NR (2006) Facing hazards and disasters: Understanding human dimensions. National Academies Press

Cutter SL (2012) Hazards vulnerability and environmental justice. Routledge

Cutter SL, Boruff BJ, Shirley WL (2003) Social vulnerability to environmental hazards. Soc Sci Q 84:242–261

Cutter SL, Emrich CT, Gall M, Reeves R (2017) Flash Flood Risk and the Paradox of Urban Development. Nat Hazards Rev 19:5017005

Cutter SL, Emrich CT, Morath DP, Dunning CM (2013) Integrating social vulnerability into federal flood risk management planning. J Flood Risk Manag 6:332–344

Cutter SL, Finch C (2008) Temporal and spatial changes in social vulnerability to natural hazards. Proc Natl Acad Sci 105:2301–2306

Dickey DA, Fuller WA (1979) Distribution of the estimators for autoregressive time series with a unit root. J Am Stat Assoc 74:427–431

Douinot A, Roux H, Garambois P-A, et al (2016) Accounting for rainfall systematic spatial variability in flash flood forecasting. J Hydrol 541:359–370

Du J, Qian L, Rui H, et al (2012) Assessing the effects of urbanization on annual runoff and flood events using an integrated hydrological modeling system for Qinhuai River basin, China. J Hydrol 464:127–139

Economic UND of (2007) World population prospects: The 2006 revision. United Nations Publications

Eftekhar B, Mohammad K, Ardebili HE, et al (2005) Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Med Inform Decis Mak 5:3

Filmer D, Pritchett L (1998) Estimating wealth effects without expenditure data—or tears. In: Policy Research Working Paper 1980, The World. Citeseer

Fischer AP, Paveglio T, Carroll M, et al (2013) Assessing Social Vulnerability to Climate Change in Human Communities near Public Forests and Grasslands: A Framework for Resource Managers and Planners. J For 111:357–365

Folke C (2006) Resilience: The emergence of a perspective for social–ecological systems analyses. Glob Environ Chang 16:253–267

Goulden ML, Bales RC (2014) Mountain runoff vulnerability to increased evapotranspiration with vegetation expansion. Proc Natl Acad Sci 111:14071–14075

Gourley JJ, Erlingis JM, Smith TM, et al (2010) Remote collection and analysis of witness reports on flash floods. J Hydrol 394:53–62. doi: https://doi.org/10.1016/j.jhydrol.2010.05.042

Gourley JJ, Flamig ZL, Vergara H, et al (2017) The FLASH Project: improving the tools for flash flood monitoring and prediction across the united states. Bull Am Meteorol

Soc 98:361–372

Hahn MB, Riederer AM, Foster SO (2009) The Livelihood Vulnerability Index: A pragmatic approach to assessing risks from climate variability and change—A case study in Mozambique. Glob Environ Chang 19:74–88

Halmstad A, Najafi MR, Moradkhani H (2013) Analysis of precipitation extremes with the assessment of regional climate models over the Willamette River Basin, USA. Hydrol Process 27:2579–2590

Hapuarachchi HAP, Wang QJ, Pagano TC (2011) A review of advances in flash flood forecasting. Hydrol Process 25:2771–2784

Hardy J, Gourley JJ, Kirstetter P-E, et al (2016) A method for probabilistic flash flood forecasting. J Hydrol 541:480–494

Hong Y, Adhikari P, Gourley JJ (2013) Flash flood. In: Encyclopedia of Natural Hazards. Springer, pp 324–325

Javelle P, Braud I, Saint-Martin C, et al (2016) Improving flash flood forecasting and warning capabilities.

Joe H (1997) Multivariate models and multivariate dependence concepts. CRC Press

Kasperson RE (2005) Ecosystems and Human Well-Being. Current State and Trends, eds Hassan RM, Scholes R, Ash N

Kauffeldt A, Wetterhall F, Pappenberger F, et al (2016) Technical review of large-scale hydrological models for implementation in operational flood forecasting schemes on continental level. Environ Model Softw 75:68–76

Kelkar U, Balachandra P, Gurtoo A (2011) Assessing Indian cities for vulnerability to climate change. In: Proceedings of the 2nd international conference on environmental science and development IPCBEE

Khajehei S, Ahmadalipour A, Moradkhani H (2018) An effective post-processing of the North American multi-model ensemble (NMME) precipitation forecasts over the

continental US. Clim Dyn 51:457–472

Khajehei S, Moradkhani H (2017) Towards an improved ensemble precipitation forecast: A probabilistic post-processing approach. J Hydrol 546:476–489

Killiches M, Kraus D, Czado C (2018) Model distances for vine copulas in high dimensions. Stat Comput 28:323–341

Klein RJT, Nicholls RJ, Thomalla F (2003) Resilience to natural hazards: How useful is this concept? Glob Environ Chang Part B Environ Hazards 5:35–45

Kleinen T, Petschel-Held G (2007) Integrated assessment of changes in flooding probabilities due to climate change. Clim Change 81:283–312

Koenker R, Hallock KF (2001) Quantile regression. J Econ Perspect 15:143–156

Kolman E, Margaliot M (2005) Are artificial neural networks white boxes? IEEE Trans Neural Networks 16:844–852

Kotzee I, Reyers B (2016) Piloting a social-ecological index for measuring flood resilience: A composite index approach. Ecol Indic 60:45–53

Kraus D, Czado C (2017) D-vine copula based quantile regression. Comput Stat Data Anal 110:1–18. doi: https://doi.org/10.1016/j.csda.2016.12.009

Kunkel KE, Karl TR, Easterling DR, et al (2013) Probable maximum precipitation and climate change. Geophys Res Lett 40:1402–1408. doi: 10.1002/grl.50334

Madadgar S, Moradkhani H (2013) A Bayesian Framework for Probabilistic Seasonal Drought Forecasting. J Hydrometeorol 14:1685–1705. doi: 10.1175/JHM-D-13-010.1

Madadgar S, Moradkhani H (2014a) Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. Water Resour Res n/a-n/a. doi: 10.1002/2014WR015965

Madadgar S, Moradkhani H (2014b) Spatio-temporal drought forecasting within Bayesian networks. J Hydrol 512:134–146. doi:

http://dx.doi.org/10.1016/j.jhydrol.2014.02.039

Madadgar S, Moradkhani H, Garen D (2014) Towards improved post-processing of hydrologic forecast ensembles. Hydrol Process 28:104–122. doi: 10.1002/hyp.9562

Maddox RA, Canova F, Hoxit LR (1980) Meteorological characteristics of flash flood events over the western United States. Mon Weather Rev 108:1866–1877

Malakar K, Mishra T (2017) Assessing socio-economic vulnerability to climate change: a city-level index-based approach. Clim Dev 9:348–363

Manly BFJ, Alberto JAN (2016) Multivariate statistical methods: a primer. Chapman and Hall/CRC

Marchi L, Borga M, Preciso E, Gaume E (2010) Characterisation of selected extreme flash floods in Europe and implications for flood risk management. J Hydrol 394:118–133

McKenzie DJ (2005) Measuring inequality with asset indicators. J Popul Econ 18:229–260

Miller JD, Kim H, Kjeldsen TR, et al (2014) Assessing the impact of urbanization on storm runoff in a peri-urban catchment using historical change in impervious cover. J Hydrol 515:59–70

Montz BE, Gruntfest E (2002) Flash flood mitigation: recommendations for research and applications. Glob Environ Chang Part B Environ Hazards 4:15–22

Najafi MR, Moradkhani H (2014) A hierarchical Bayesian approach for the analysis of climate change impact on runoff extremes. Hydrol Process 28:6292–6308

Najafi MR, Moradkhani H (2015) Ensemble combination of seasonal streamflow forecasts. J Hydrol Eng 21:4015043

Nasrabadi NM (2007) Pattern recognition and machine learning. J Electron Imaging 16:49901

Nelsen RB (1999) An introduction to copulas. Springer

Norbiato D, Borga M, Degli Esposti S, et al (2008) Flash flood warning based on rainfall thresholds and soil moisture conditions: An assessment for gauged and ungauged basins. J Hydrol 362:274–290. doi: https://doi.org/10.1016/j.jhydrol.2008.08.023

Nyamundanda G, Brennan L, Gormley IC (2010) Probabilistic principal component analysis for metabolomic data. BMC Bioinformatics 11:571. doi: 10.1186/1471-2105-11-571

Ogden FL, Sharif HO, Senarath SUS, et al (2000) Hydrologic analysis of the Fort Collins, Colorado, flash flood of 1997. J Hydrol 228:82–100. doi: https://doi.org/10.1016/S0022-1694(00)00146-3

Rana A, Moradkhani H, Qin Y (2017) Understanding the joint behavior of temperature and precipitation for climate change impact studies. Theor Appl Climatol 129:321–339

Reza Najafi M, Moradkhani H (2013) Analysis of runoff extremes using spatial hierarchical Bayesian modeling. Water Resour Res 49:6656–6670

Saha S, Moorthi S, Pan H-L, et al (2010) The NCEP climate forecast system reanalysis. Bull Am Meteorol Soc 91:1015–1057

Saha S, Moorthi S, Wu X, et al (2014) The NCEP climate forecast system version 2. J Clim 27:2185–2208

Saharia M, Kirstetter P-E, Vergara H, et al (2017b) Mapping flash flood severity in the United States. J Hydrometeorol 18:397–411

Saharia M, Kirstetter P-E, Vergara H, et al (2017a) Characterization of floods in the United States. J Hydrol 548:524–535

Sangati M, Borga M, Rabuffetti D, Bechini R (2009) Influence of rainfall and soil properties spatial aggregation on extreme flash flood response modelling: an evaluation based on the Sesia river basin, North Western Italy. Adv Water Resour 32:1090–1106

Schmidtlein MC, Deutsch RC, Piegorsch WW, Cutter SL (2008) A sensitivity analysis of the social vulnerability index. Risk Anal An Int J 28:1099–1114

Seaber PR, Kapinos FP, Knapp GL (1987) Hydrologic unit maps

Shirley WL, Boruff BJ, Cutter SL (2012) Social vulnerability to environmental hazards. In: Hazards Vulnerability and Environmental Justice. Routledge, pp 143–160

Singh VP, Woolhiser DA (2002) Mathematical modeling of watershed hydrology. J Hydrol Eng 7:270–292

Sklar M (1959) Fonctions de répartition à n dimensions et leurs marges. Université Paris 8

Solomatine DP, Ostfeld A (2008) Data-driven modelling: some past experiences and new approaches. J hydroinformatics 10:3

Solomatine DP, PRICE RK (2004) Innovative approaches to flood forecasting using data driven and hybrid modelling. In: Hydroinformatics: (In 2 Volumes, with CD-ROM). World Scientific, pp 1639–1646

Špitalar M, Gourley JJ, Lutoff C, et al (2014) Analysis of flash flood parameters and human impacts in the US from 2006 to 2012. J Hydrol 519:863–870

Stephens MA (1974) EDF statistics for goodness of fit and some comparisons. J Am Stat Assoc 69:730–737

Teegavarapu RS V (2019) Changes and Trends in Precipitation Extremes and Characteristics: Links to Climate Variability and Change. In: Trends and Changes in Hydroclimatic Variables. Elsevier, pp 91–148

Terti G, Ruin I, Anquetin S, Gourley JJ (2015) Dynamic vulnerability factors for impact-based flash flood prediction. Nat Hazards 79:1481–1497

Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. J R Stat Soc Ser B (Statistical Methodol 61:611–622

Trenberth KE, Dai A, Rasmussen RM, Parsons DB (2003) The changing character of

precipitation. Bull Am Meteorol Soc 84:1205–1217

Tu J V (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 49:1225–1231

Turner BL, Kasperson RE, Matson PA, et al (2003) A framework for vulnerability analysis in sustainability science. Proc Natl Acad Sci 100:8074–8079

Vanuytrecht E, Van Mechelen C, Van Meerbeek K, et al (2014) Runoff and vegetation stress of green roofs under different climate change scenarios. Landsc Urban Plan 122:68–77

Verhoest NEC, Van Den Berg MJ, Martens B, et al (2015) Copula-based downscaling of coarse-scale soil moisture observations with implicit bias correction. IEEE Trans Geosci Remote Sens 53:3507–3521. doi: 10.1109/TGRS.2014.2378913

Vyas S, Kumaranayake L (2006) Constructing socio-economic status indices: how to use principal components analysis. Health Policy Plan 21:459–468

Wanders N, Karssenberg D, Roo A de, et al (2014) The suitability of remotely sensed soil moisture for improving operational flood forecasting. Hydrol Earth Syst Sci 18:2343–2357

Wang W, Chau K, Xu D, et al (2017) The annual maximum flood peak discharge forecasting using Hermite projection pursuit regression with SSO and LS method. Water Resour Manag 31:461–477

Wigtil G, Hammer RB, Kline JD, et al (2016) Places where wildfire potential and social vulnerability coincide in the coterminous United States. Int J Wildl fire 25:896–908

Wu CL, Chau KW, Li YS (2009) Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. Water Resour Res 45:

Yan H, Moradkhani H (2016) Toward more robust extreme flood prediction by Bayesian hierarchical and multimodeling. Nat Hazards 81:203–225

Yilmaz I (2010) Comparison of landslide susceptibility mapping methodologies for

Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine. Environ Earth Sci 61:821–836

Zarekarizi M, Rana A, Moradkhani H (2018) Precipitation extremes and their relation to climatic indices in the Pacific Northwest USA. Clim Dyn 50:4519–4537

Zhang J, Howard K, Langston C, et al (2016) Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. Bull Am Meteorol Soc 97:621–638

Zhang L, Singh VP (2014) Trivariate flood frequency analysis using discharge time series with possible different lengths: Cuyahoga river case study. J Hydrol Eng 19:5014012

Zhao T, Minsker B, Salas F, et al (2018) Statistical and Hybrid Methods Implemented in a Web Application for Predicting Reservoir Inflows during Flood Events. JAWRA J Am Water Resour Assoc 54:69–89